Stakeholder report M1

Mikkel Bak Lyngø

24 sep 2021

# Background and objectives for the Group Assignment

In this assignment, we had the opportunity to use newly acquired abilities within Machine Learning to investigate a given problem statement. Through the M1 course, we have practised unsupervised machine learning techniques for dimensionality reduction and clustering to discover relationships between features and groupings of observations. In addition, we have used supervised machine learning for regression and classification problems, where we created models to predict an outcome of interest given some input features. Based on this, we will try to carry out an analysis which contains elements of data manipulation, exploration, unsupervised and supervised machine learning.

# Problem statement

Use machine learning techniques to investigate which determinants are decisive for a startup's success. A startup is a young company or project founded by one or more entrepreneurs to develop a unique product or service and bring it to market. There has been an exponential growth in startups over the past few years and startups play a major role in economic growth. They bring new ideas, spur innovation, create employment and thereby moving the economy, which is why it is interesting to perform an analysis on these startup's.

# Data acquisition

Given the problem statement we then choose and obtain a dataset which we consider interesting and appropriate for this analysis. Through Kaggle (which contains more than 50,000 public datasets) we were able to find a relevant dataset. The data contains 48 columns/features, which describes different industry trends, investment insights and individual company information.

# Results

We will continuously refer to the current steps in the analysis, using ("1"), which is also associated with all parts of the code version of the assignment.

We start by loading all the relevant packages (1) which are used to later perform various functions in R. Then we load in our dataset (2) and in the same step we convert some date-variables to numerical values as these variables will be included in the pca later on. In the data cleaning process we first take a glimpse of the data (3) to get a better look at it. We start by removing columns we dont know what describes as they are not described when loading the dataset. Also some of the columns show the same thing and therefore removed as well (4) . To look how wholesome the data is we skim it and observe some NA´s which we want to get rid of. In the variable "closed\_at" we replace NA´s with "still going", which means that the startup is not closed (5, 6, 7). Then we divide the variable "first\_milestone" into intervals (8), as we want to investigate whether the duration of the first milestone has significance for the other variables of interest. In addition, we limit both "state\_code" and "category\_code" to preserve the most interesting outcomes and let the others be defined as "other". This completes our data cleaning process.

Moving on to the exploratory data analysis (EDA) we choose the variables that we find most relevant. This includes "funding\_total\_usd", "funding\_rounds", "relationships","age\_first\_funding\_year", "milestones" and "status".

# Conclusions