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 skim the data (3) to see how wholesome it is. We start by removing columns that are undefined as they are not described when loading the dataset. Some of the columns show the same thing and are therefore removed as well. We also remove variables like "longtitude" and "latitude", which could be fun to work with later on, but doesnt seem to bring much insight to the ongoing analysis (4).

By removing those we now have our dataset containing 33 numerical variables. We observe that 22 of those are dummy variables, which means that the variable only have 2 outcomes. It is possible to use dummy variables when trying to perform cluster analysis and see if our clusters are separated by selected dummy variables. However our main objective is to try and separate by the "status" column, which contains the outcomes: "closed" or "acquired". We therefore want to find out, which variables seems to be correlated with the "status" column, starting with the jobtype related dummies, the state related dummies and lastly the three remaining dummies.

Selecting the three dummy categories and plotting those separately against the variable "status" we get our 3 plots for correlation: The plot of the job type related dummies shows that there is almost no differences in how many firm closes and how many who becomes acquired according to job type. Looking at the state related dummies, the plot shows that a bit more firms from New York, Texas and California are being acquired than from the rest of the United States, which could indicate that they could be useful variables to use in our analysis. On the other hand you could argue, that the differences aren't substantial which would make them less influential in the analysis. Lastly in the plot of the three remaining dummy variables it is clear to see that whether or not the firm has been a top 500 upcoming firm substantially changes the chance of closing or not. Whether they have VC or angel investors doesn't seem to influence the outcome in a big way. Given the results from our three plots we drop the "state" and "job" related variables plus the VC and angel dummies, however we keep "is\_top500" and then we skim the data once again.

Still there are 8 numerical variables as well as 3 date variables we haven't taken a closer look at yet, which naturally takes us to the part where we determine how those variables seems to be correlated with the "status" column. By performing a plot containing "relationships", "milestones", "funding\_total\_usd", "funding\_rounds", "closed\_at", both "first and last milestone\_year" and "last\_funding\_year" we observe that "closed\_at" will not be used because it contains almost 600 missing observations which is close to 66% of all observations. Futhermore, "founded\_at" will not be used either simply because there doesnt seem to be that big of a difference in the density plots. Using both "first and last milestone\_year" might be irrelevant because they somehow nearly show the same thing, which is why we dont use "age\_last\_milestone\_year". The same goes for "first and last funding\_at" which doesn't show any significance either.

After the removal of selected variables we are left with only our "status" variable and our 7 numerical variables of interest. For those variables that remains we check for missing values. We observe that "age\_first\_milestone\_year" has 152 missing values. However the problem with this variable is, that if we remove those rows, we lose 16.4% of our observations, and we cant just replace the NA's with something else - like zero - because that would only manipulate our data. The solution we have come up with is to divide the variable into intervals, as we want to investigate whether the duration to the first milestone has significance in terms of a startup getting acquired. We pick the interval zero to three years and we drop the rest, which means that startup achieved their first milestone when they were between 0 and 3 years old. This concludes our data preparation process.

Moving on to the exploratory data analysis (EDA) we now have our variables of interest:

names = colnames(data)  
description = c("numeric, age at last/latest funding", "numeric, number of relationships", "numeric, number of funding rounds", "numeric, the total funding in USD", "numeric, number og milestones", "dummy variable, has been a top 500 start up", "character string, acquired or closed", "reach their first milestone at the age of 0-3 years")  
  
table = cbind(names, description); table

## description   
## [1,] "numeric, age at last/latest funding"   
## [2,] "numeric, number of relationships"   
## [3,] "numeric, number of funding rounds"   
## [4,] "numeric, the total funding in USD"   
## [5,] "numeric, number og milestones"   
## [6,] "dummy variable, has been a top 500 start up"   
## [7,] "character string, acquired or closed"   
## [8,] "reach their first milestone at the age of 0-3 years"

We start by performing dimensionalty reduction on our data, however we need to know whether our data should be scaled before we can do so. We observe that our data has to be scaled in order to peform PCA. Namely "funding\_total\_usd" seems much higher. We also have some dummy variables which only take a value of either 0 or 1, so these will also be on a different scale than the rest. We scale the data by setting the argument scale.unit to TRUE and then we run the PCA. Further we can only run our PCA on numeric variables and therefore we use the data subset called "data\_num".

# Conclusions