1 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.

2 Problem statement

Use machine learning techniques to investigate whether we can make a model predict if a startup will close or be acquired. When doing so we also want to know 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.

3 Data acquisition

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

After tidying the data eight variables are chosen, which include seven numeric variables(**age-last-funding-year, relationships, funding-rounds, milestones, is-top500, total** and **is-0:3**), two of them being binary(**is-top500** and **is-0:3**), and one character string(**status**). These will now be described.

Age-last-funding-year describes the age of a company in years when it got its last or latest funding. **Relationships** shows how many relationships a startup has. For examp-

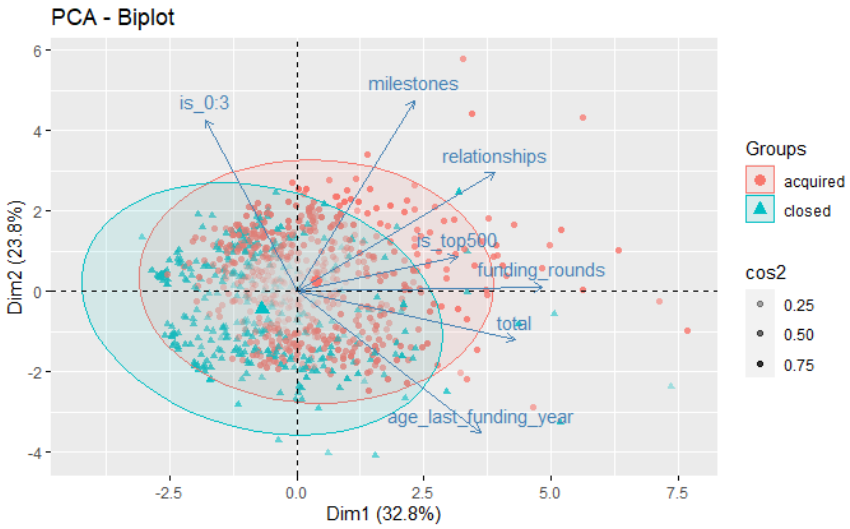
le a start up can have relationships with accountants, investors, vendors, mentors, etc. **Funding-rounds** is the number of rounds a firm has recieved funding. **Milestones** tracks startups' progress as a startup grows and implements their plan. **Is-top500** is a binary variable being 1 if a company at some point has been on the top 500 startup list otherwise 0.**Total** describes the total amount of funding a company has recieved in 1000 USD. **Is-0:3** is a selfmade binary variable being 1 if a company reach it's first milestone within 0-3 years of being active otherwise 0. **Status** is a character variable which describes the status of a company either being "closed"if the company had shut down and being "acquired"if the company had been acquired.

4 Results

4.1 Unsupervised Machine Learning

The main objective is to try and separate the data into different clusters and see if the data is clustered by status, or other categorical variables from the data. First step will be to perform dimensionality reduction on the data. It can be observed that the data has to be scaled in order to perform PCA. Once the data has been scaled it is possible to pick the optimal number of dimensions by using a screeplot or eigenvalues. When looking at the eigenvalues the rule of thumb is to pick the dimensions with an eigenvalue >=1, which in this case gives us two dimensions.

From figure 1 it can be observed that the x-axis split the seven variables into two groups, where the ones involving funding or money are below or on the x-axis. Funding-rounds lies on the x-axis which indicates that it is mostly explained by the first dimension. The remaining variables all lie above



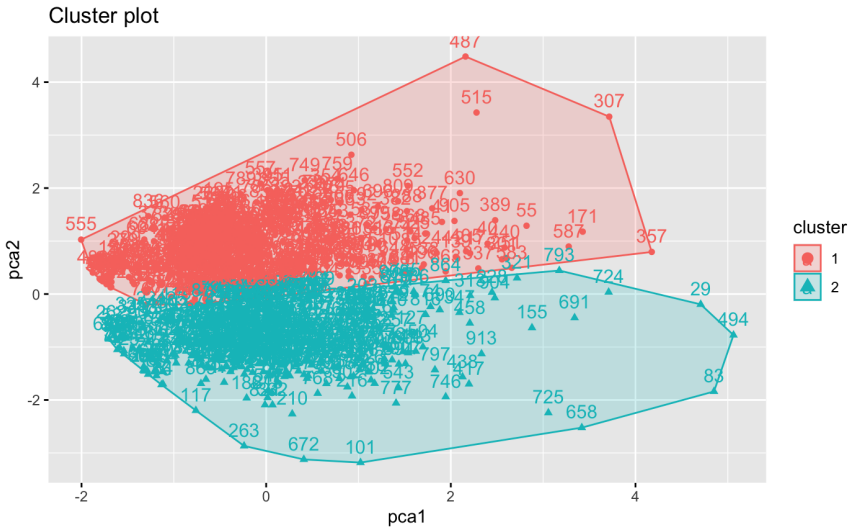
Figur 1: PCA

the x-axis. It can be said about the variables, that the less opaque the dot the higher the "cos2"and thereby a higher representation of the variable in the principal components used. However the conclusion would be that the PCA doesn't really separate the startup´s by status.

4.1.1 Clustering using K-means

After performing both K-means clustering and hierarchical clustering on both the entire dataset and on the PCA data the best result was given by the k-means clustering on PCA data which can be seen in figure 2.

The screeplot suggests that we should use three clusters, but as we want to cluster by the "status-variable, we use two clusters. Clustering with the PCA data seems to make two clusters which are nicely separated. To see if the clusters are separated by status we take a look at table 1. By using the dimensionality reduced data it seems like the startups haven't been separated by status that well.



Figur 2: PCA

Tabel of K-means clustering on PCA-data		
	Acquired	Closed:
Cluster 1	352	144
Cluster 2	241	181

Tabel 1: Single equation ARDL-Bounds LR-estimates

4.2 Supervised Machine learning

This section will give an explanation of the results of making a model that can predict if a startup will close or be acquired using supervised Machine Learning.

In this section five models are made, which includes a logistic regression model, a decision tree model, a random forest model, a K-nearest neighboors model and lastly a XGBoost model. Before running the data on the five models the data has been split into a training and test data set, using 75% of the data as training data. The models are trained on data created from crossvalidation. The results of each model will be discussed in the next section. Before we run the models on the data, we create a recipe which will log transform right scewed data, center all the numeric variables to mean equal to 0, scale all numeric varaibles to a standard deviation of 1. When we fit the models we use hyperparameter tuning for the random forest model and the XGBoost model. The goal is to not overfit the models so the variance error will not be to large, but still keep the model flexible to minimize the bias error.

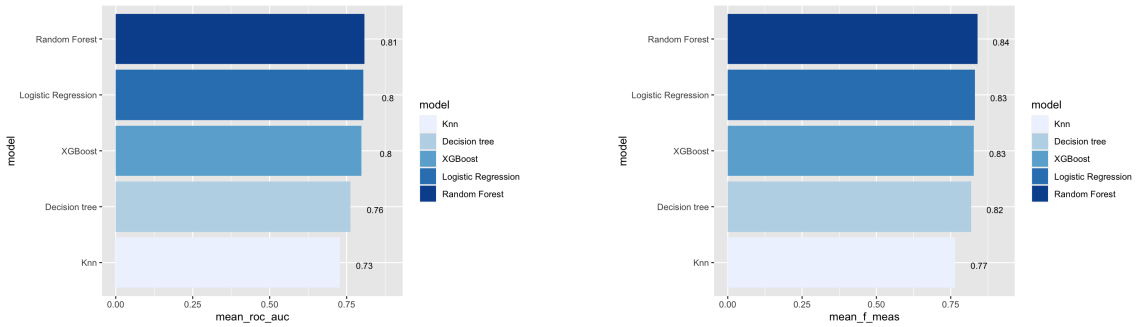
4.2.1 Model comparison

By looking at figure 3 We can see we get the highest mean ROC under the curve value using the Random Forest model. The same model scores the highest value using the mean f-measure. This measure is a weighted average of two other variables: Recall and precision. Recall is the amount of true positives in the model divided by the number of true positives **and** the amount of false negatives. Thereby a low recall value will indicate a large amount of false negatives. Precision is a number formed by taking the amount of true positives in the model divided by the amount of true positives **and** true negatives. After also looking at the confusion matrices of the different models, we conclude that the best model to predict if startups will close or be acquired is the Random forest model. The next step will be to analyze how the chosen model performs on the testing data.

4.2.2 Random forest model on test data

The model is now used to predict the status only given the test data. By looking at figure 4 the confusion matrix shows that the model predicts that startups will be acquaried

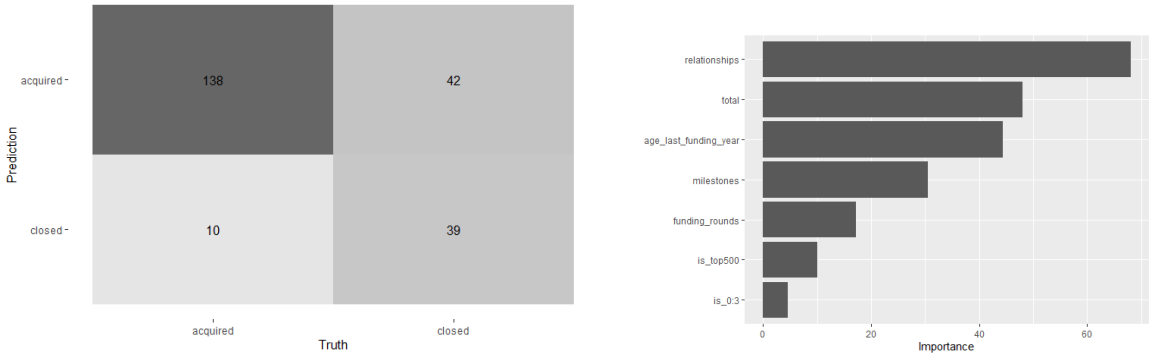
correctly 138 times (True positives), but predictes that the startups will be acquired incorrectly 42 times (False positive). The same can be seen for closed companies showing 10 False negative and 39 True positives. This shows almost the same results as running it on training data.



(a) Mean_ROC_AUC **(b)** Mean_f_meas

Figur 3: Model Comparison plot

From figure 4 the variables' importance in the model is shown where it can be observed that the most important factor for predicting if a startup will close or be acquired is the amount of relationships, and the total funding the startup has acquired.



(a) Confusion matrix **(b)** label 2

Figur 4: Random forest on test data

All code and elaborations can be found in the attached pdf file or alternatively by following this Google Colab link.