Investigating the Potential of ML in Detecting Attractive Investments from Result and Balance Sheet Data

II2202 oral exam

Team members: Petrus Oskarsson and Robin De Groot

Topic: Investigating the Potential of ML in Determing Attr. Inv. only training on Result and Balance Sheet Data

Introduction

Background and rationale

- **Previous interesting research has been done** on how machine learning can be applied to detect attractive investments
- However, these models are often built on data with **restricted access**
- Hence, it would be beneficial to **explore the potential of a supervised ML model only based on easily accessible data**
  - **-** E.g. financial figures from **result and balance sheet data**
- **Sweden is a good case study** for this purpose as companies are obliged to report financial figures from result and balance sheet to 'Skatteverket'[1]

Goals, hypothesis, research question and expected outcomes

- Goal: to **explore the potential of supervised machine learning models based on this data**
- Hypothesis: Even though other information and soft variables usually also go into a investment decision, it **can still help** an investment decision
- Research question: Can a supervised ML model only trained on result and balance sheet data, **be valuable to determine attractive invest.?**
  - Which ML model is **most accurate** on this task?
  - What are **important factors** to explain the decision making of investors currently?
- Expected outcomes: **comparison** of model performance

Source: 1. Skatteverket web site

Q: Can a supervised ML model predict attr. inv. only trained on result and balance sheet data? Driving factors?

Introduction

Research questions

- Can a supervised ML model **trained exclusively on results and balance sheet data** be useful to determine attractive investment?

- Which of selected **ML model is most accurate** on this task?

- What are **important factors** to explain investment decision making?

- Introduction
- **Executive summary of results so far**
- Data exploration
- Important factors
- Evaluated data sets
- Parameter settings
- Model comparison and selection
- Final performance
- PCA modelling
- Discussion and future work
- Appendix

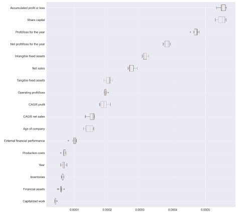
Some driving factors have been identified and the Gradient Boosting Classifier showcased the best performance with AUC 0,5777

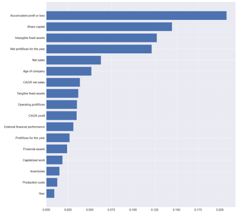
Executive summary of results so far

Illustrative images

Driving factors

- **‘Accumulated profit or loss’** seem to be a especially strong predictor followed by **’share capital’ and others**





Best performing model

- The Gradient Boosting Classifier was the best performing model – the data set with 24 variables could be the optimal one to use

Classifier	Cross validation AUC scores		
	Dataset 1	Dataset 2	Dataset 3
Decision Tree Classifier	0,9183973	0,9138278	0,915044
Random Forest Classifier	0,9392665	0,940013	0,9376575
AdaBoost Classifier	0,9363752	0,9354571	0,9301463
Gradient Boosting Classifier	**0,9450325**	**0,9457589**	**0,9432907**
Support Vector Classifier	0,5414041	0,465179	0,4539837

Final performance

- The final AUC performance score is 0,5777
- Only a bit better than random

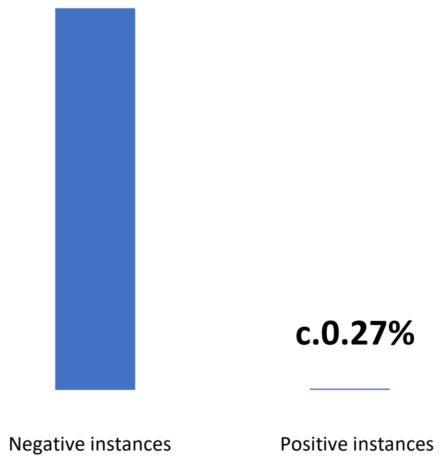
	AUC score on validation data	
	Dataset 1	Dataset 2
Gradient Boosting Classifier	0,5770	0,5777

- Introduction
- Executive summary of results so far
- **Data exploration**
- Important factors
- Evaluated data sets
- Parameter settings
- Model comparison and selection
- Final performance
- PCA modelling
- Discussion and future work
- Appendix

We are dealing with an imbalanced detection problem, 0.27% of positive instances – AUC will be used as metrics

Data exploration

Imbalanced data set

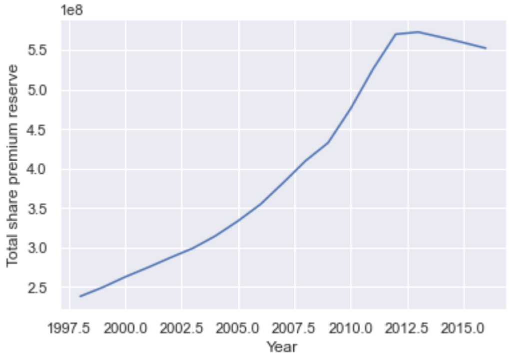


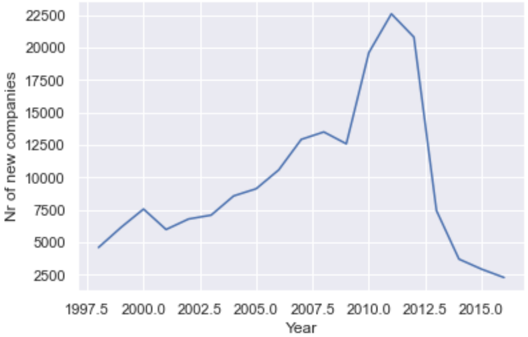
- Only **0.27%** of poitive instances
- To cope with this imbalance nature of the data set, **AUC will be devoted as the metrics**

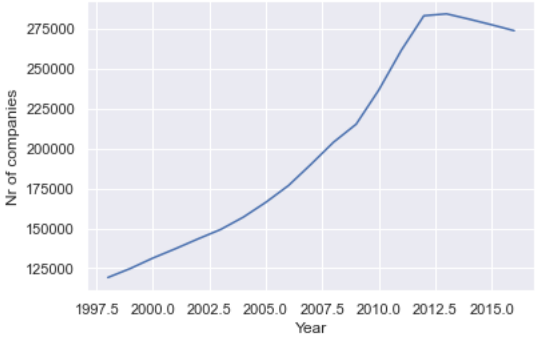
To cope with unexpected behaviours and to enable model running, only years up until 2012 is used

Data exploration

Year over year analysis







- Some unexpected behavior was seen in the yby analysys, **a downward trend 2012 and onwards**
- After validation with industry experts, it was concluded **this is not expected behavior**
- There was also a need to partition the data set to enable model running, a pragmatic solution was made where **years 2012 and onwards were excluded**

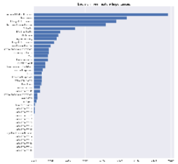
- Introduction
- Executive summary of results so far
- Data exploration
- **Important factors**
- Evaluated data sets
- Parameter settings
- Model comparison and selection
- Final performance
- PCA modelling
- Discussion and future work
- Appendix

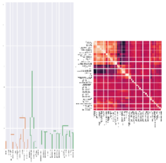
Some initial feature reduction has been performed with the help of importance- and hierarchical cluster analysis

Important factors driving investment decisions – feature reduction

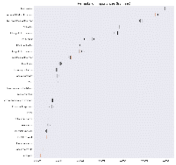
Illustrative

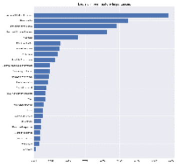
40 features

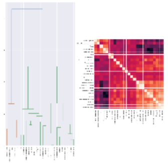




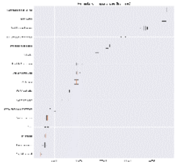
24 features

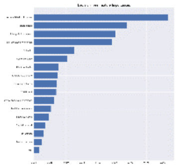


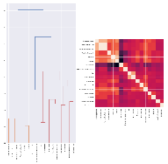




17 features



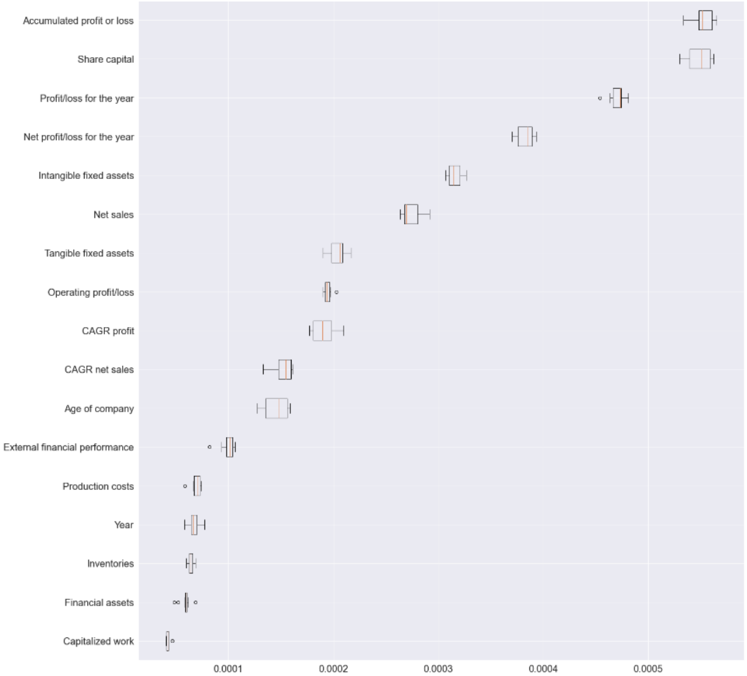




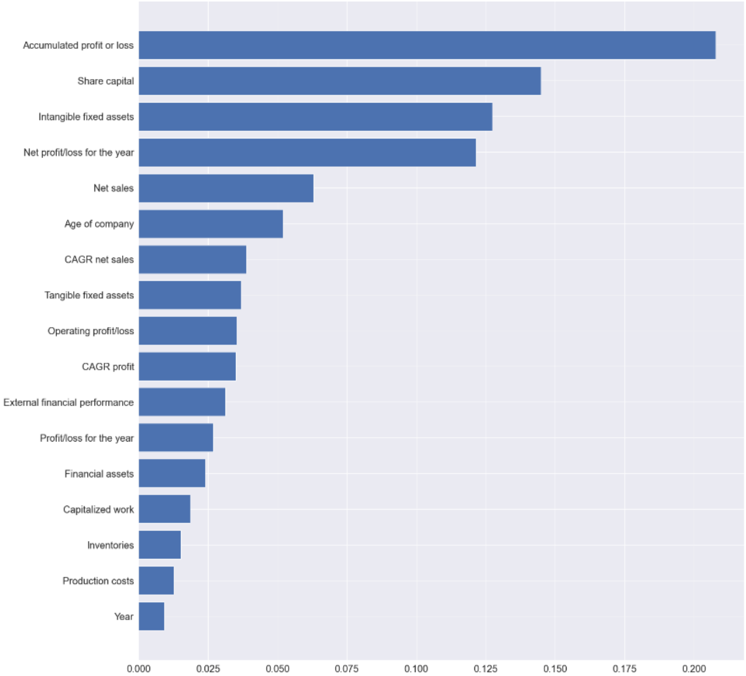
‘Accumulated profit or loss’ seem to be a especially strong predictor followed by ’share capital’ and others

Important factors driving investment decisions

Permutation feature importance[1]



Decision tree feature importance[2]



Topmost imp. feat.

- **Accumulated profit or loss**
- Share capital
- Net profit/loss for the year
- Intangible fixed assets
- Net Sales
- Profit/loss for the year
- Tangible fixed assets
- Operating profit/loss
- CAGR net sales
- CAGR profit

1) Measure impact of prediction by shuffling each feature column, 2) Based on ”Gini-importance”

- Introduction
- Executive summary of results so far
- Data exploration
- Important factors
- **Evaluated data sets**
- Parameter settings
- Model comparison and selection
- Final performance
- PCA modelling
- Discussion and future work
- Appendix

Three different data sets have been evaluated to compare performance across models and feature sets

Evaluated data sets

Data sets

- Three different data sets will be evaluated to compare performance across models and feature sets
- The feature reduction to set 24 resp. 17 features have been conducted based on the feature importance and hierarchical clustering

- The data sets that have been evaluated are the following:

1. The first data set has **40 features**
2. The second one has **24 features**
3. And the third, **17 features**

- Introduction
- Executive summary of results so far
- Data exploration
- Important factors
- Evaluated data sets
- **Parameter settings**
- Model comparison and selection
- Final performance
- PCA modelling
- Discussion and future work
- Appendix

Some extensive, especially time sufficient, parameter tuning have been performed

Parameter settings

Hyper parameter tuning

- Some extensive, especially time sufficient, parameter tuning have been made
- 10 different parameter sets were tested per model, each run 3 times with cross validation
- Due to the time intensive process models have only been hyper-tuned on the first data set for as per today

Classifier	Parameter	Value	Parameter	Value
Decision Tree Classifier	splitter	best	min_samples_split	11
Random Forest Classifier	n_estimators	225	min_samples_split	5
Nearest Neighbour Classifier	weights	distance	p	2
AdaBoost Classifier	n_estimators	490	learning_rate	0.5
XGBoost Classifier	reg_lambda	10	n_estimators	310

Classifier	Parameter	Value	Parameter	Value
Decision Tree Classifier	max_depth	7	criterion	entropy
Random Forest Classifier	max_features	sqrt	max_depth	9
Nearest Neighbour Classifier	n_neighbors	6	algorithm	ball_tree
AdaBoost Classifier				
XGBoost Classifier	min_child_weight	7	max_depth	20

Classifier	Parameter	Value	Parameter	Value
Decision Tree Classifier				
Random Forest Classifier	criterion	entropy		
Nearest Neighbour Classifier				
AdaBoost Classifier				
XGBoost Classifier	learning_rate	0.1	gamma	1

- Introduction
- Executive summary of results so far
- Data exploration
- Important factors
- Evaluated data sets
- Parameter settings
- **Model comparison and selection**
- Final performance
- PCA modelling
- Discussion and future work
- Appendix

The Gradient Boosting Classifier was the best performing model – data set 2 could be the optimal one to use

Model comparison and selection

Performance across feature sets and models

- **Selected model is the Gradient Boosting Classifier**, from the XGBoost library - **outperformed the other models**
- We will **discard "Dataset 3"** as the performance is lower and did not see a good enough improvement on the run-time
- However, we find the **difference between data set 1 and 2 is hard to judge as** the deltas could occur due to chance
- Apparently, dropping the 16 variables from data set 1 to 2, **does not seem to have any larger impact**

Cross-validation on training data

Classifier	Cross validation AUC scores		
	Dataset 1	Dataset 2	Dataset 3
Decision Tree Classifier	0,9183973	0,9138278	0,915044
Random Forest Classifier	0,9392665	0,940013	0,9376575
AdaBoost Classifier	0,9363752	0,9354571	0,9301463
Gradient Boosting Classifier	**0,9450325**	**0,9457589**	**0,9432907**
Support Vector Classifier	0,5414041	0,465179	0,4539837

# of variables: 40 var. 24 var. 17 var.

-16 var. -7 var.

- Introduction
- Executive summary of results so far
- Data exploration
- Important factors
- Evaluated data sets
- Parameter settings
- Model comparison and selection
- **Final performance**
- PCA modelling
- Discussion and future work
- Appendix

Performance on AUC is 0,5777 - only a bit better than random and no major difference between data set 1 & 2

Final performance

Gradient Boosting Classifier performance

- The performance is a bit better than random
- Performance difference between data set 1 and 2 is very small
- One could pereferably go with data set 2 as it contains fewer features with motivation that the dropped ones did not add much information

AUC performance on validation data

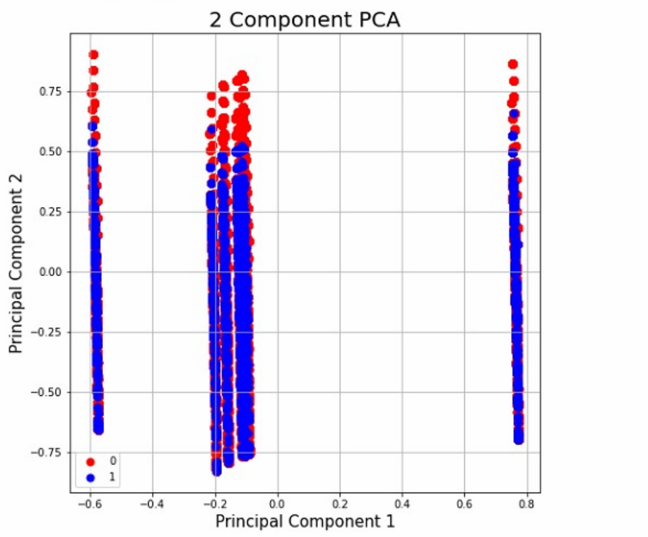
AUC score on validation data		
	Dataset 1	Dataset 2
Gradient Boosting Classifier	0,5770	0,5777

- Introduction
- Executive summary of results so far
- Data exploration
- Important factors
- Evaluated data sets
- Parameter settings
- Model comparison and selection
- Final performance
- **PCA modelling**
- Discussion and future work
- Appendix

PCA modelling indicates that it is difficult to separate the data

PCA modelling

PCA visuals



- 2 Dimensional PCA on training set
- Should verify results, but classes overlap significantly
- Not all variance explained by 2 components, still possible explanation of limited performance

- Introduction
- Executive summary of results so far
- Data exploration
- Important factors
- Evaluated data sets
- Parameter settings
- Model comparison and selection
- Final performance
- PCA modelling
- **Discussion and future work**
- Appendix

The predictor is only useful to a limited extent. Whether that depends on limitations is hard to say, however, it opens up to some other interesting studies

Discussion and future work

Research limitations

- The **choice of proxy** for an ”attractive investment” is probably the most major limitation

- **Authors’ financial knowledge** is limited

- **Training/test split was done after feature selection**

- **Many rows were dropped** due to a large number of missing values

- **Excluding the years after 2012** could potentially limit the ”relevance” of the models

- **Model/dataset choice** just by CV on training set

Discussion points

- Performs slightly better than random and could indicate that there are **some useful information** that goes into the model

- However, even though some indication, but **only useful to a very limited extent**

- Potentially the high imbalance nature of the data makes it a **difficult problem to learn**

- If a model could only **slightly improves** investors ability to determine attractive invest., it **would be worth a lot**

Future work

- Similar study with a crossfunctional team with both **analytical skills** and **accounting know-hows**

- **Investigate other possible proxies** for ”a received an investment”

- Look into the potential of **”man and machine” combination** for this task
  - Using invester judgement combined with the mode
  - Improve from the performance level of an investor

An experimental study with an inductive and reductionist approach – equity column is used as proxy

Appendix A: Research methodology

Team members: Petrus Oskarsson and Robin De Groot

Tasks and research methods

> According to the reductionist approach, the research is broken down into tasks
  > First, data exploration is necessary
  > Interesting variables for the models should be determined
  > Feature engineering will most likely be necessary
  > Machine learning models will be tested on the data
  > Insights will be evaluated and discussed

> The research is an experimental study with and an inductive approach – Supervised ML models are be applied to the observed dataset to draw potential conclusions

Method discussion

> Increase equity column as a proxy for investment received

> A hard assumption in this research is that investors would have a high success rate in finding attractive investments

> Thus, if a company saw an increase in their equity column, ergo received an investment, it will be seen as an attractive investment

> Of course this is not always the case, as investors also make unsuccessful investments

Inspired by earlier papers and industry use cases, the research will leverage some well know ML methods

Appendix B: Theory and literature review

Team members: Petrus Oskarsson and Robin De Groot

ML in investment decisions

> Currently, most investors are not making use of ML technology yet
> Previous papers on area are limited but there are some researchers and companies looking into the possibilities
  > E.g. Arroyo et al. (2020) who found that investors currently do not have access to tools that allow them to reduce risk and uncertainty enough
  > E.g. Van Witteloostuijn and Kolkman (2019) who created a prediction model to estimate company growth
  > Example of industry use cases are Hone Capital[2] and EQT Motherbrain[3]

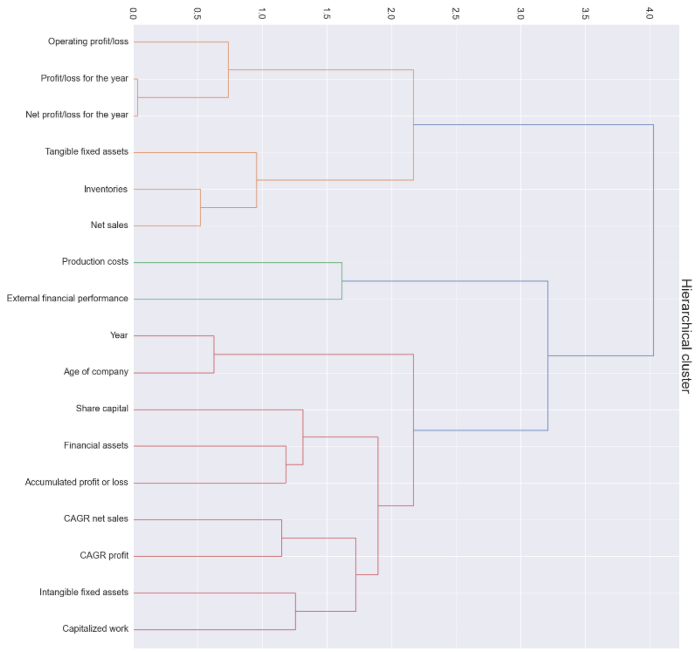
ML methods

> Multiple ML methods will be applied: decision tree, random forest, gradient boosting tree, K-nearest neighbor, support vector machine, and artificial neural network

Appendix C

Finally, w. 17 features, ‘Profit for the year’ and ‘net profit of the year’ show high collinearity

Important factors driving investment decisions



- Feature clustering show high collinearity between ‘Profit for the year’ and ‘net profit of the year’

A successful algorithm can help economic growth, contribute to innovation and help reduce inequalities

Appendix C: Key ethical and sustainability issues in context of Envision 2030

Goal #8 – help economic growth

- A successful prediction algorithm can potentially spot companies that have not received investment but in fact qualify for it.
- This way, more companies with high economic potential would be invested in, which provides increased economic growth.

Goal #9 – contribute to innovation

- Hypothetically, innovative firms should also be classified as ‘attractive investments’ and if a model can help to invest in these, it can contribute to overall growth
  - With a successful commercialized algorithm, investors could potentially process a wider scope of investment opportunities and allocate their resources to the most innovative firms

Goal #10 – reduce inequalities

- If a platform based on machine learning would be commercialized, potential inequality issues can occur if the training data is biased
  - For instance, if historic data on investments tend to be biased towards a certain ethnicity or gender, the risk is that the model too will be biased
  - The data set we have does not contain such data, however, one might add extensions that do