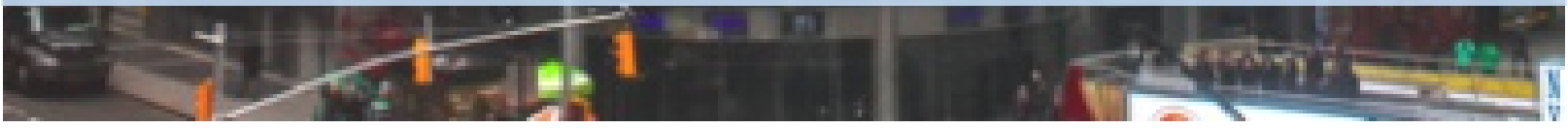
Bankruptcy Modelling

- a Machine Learning Approach

Sebastian Pedersen, Lasse Petersen, Simon Harmat & Nikolaj Thams



BUSINESS ANALYTICS CHALLENGE 2017

DANSKEBANK.DK/BAC2017

The work within this paper was concluded by students Sebastian Pedersen (stud.scient.oecon), Lasse Petersen (stud.stat), Simon Harmat (stud.polit) and Nikolaj Thams (stud.scient.oecon), all at University of Copenhagen. Ownership of this paper is shared between the those students and the BAC2017 organisers.

The paper is written with Danske Bank in mind as the receiver, and one will often find phrases like "We recommend Danske Bank to …". None of the conclusions are however limited to Danske Bank, and anyone with an interest in the default of Danish companies, would find the model applicable.

A special thanks is to be given to Prof. Niels Richard Hansen for providing statistical guidance.

JUNE 2017

Contents

1. Introduction

Through recent years, we have seen company status change suddenly and
corporations have gone into distress while other companies escalate in gro
crucial from both a society and a banking point of view that we can be proac

This paper is an answer to the question stated in the BAC2017 problem: Can we predict a company's future? We answer this by focusing on predicting company defaults one year ahead.

To do this we build a framework which extracts and refines publicly available Danish annual reports and merge this with a variety of other sources. We then utilize modern techniques from statistics and machine learning, combined with financial theory, to build models able to predict company default one year ahead. An ensemble model based on a logistic regression, a neural network, a random forest and a support vector machine using the method of stacking delivers interpretable probabilities of default, time-consistency and a stable AUC of 82%-84% over time. Using our model we are able to detect 51.8% of the defaulted companies with only a 10% false positive rate on the non-defaulted companies.

Conclusively, we suggest what value this will have to Danske Bank, and how to make further applications of our findings. To showcase this we develop a model, which shows that Danske Bank can increase their profit on business loans with 1.3 % by applying a risk-weighted process for giving loans. The profit increase is a consequence of the fact that the reward of predicting some defaulting companies is larger than the loss of refusing loans to surviving companies.

Furthermore we suggest applications that could improve Danske Bank's market position as a leading technology based bank. For example we suggest a customer benchmarking tool, in which Danske Bank would be able to swiftly deliver market leading insights to all Danske Bank's business clients. Based on our model another suggestion is the basis for a data-driven client advising tool that can suggest concrete improvements, that will lower customer defaulting probability, which will both benefit the customers and Danske Bank.

By applying a strict methodical discipline, we deliver a framework which is as stable, time independent and easily implementable. At the same time we are hoping that this paper will shine light on previously untested procedures, and the ambition is to provide inspiration for anyone with an interest in data treatment, statistical models and credit risk estimation. Enjoy the reading!

```
defaultData <- read.csv("1 CVR Udtræk/CVR_m_branche.csv", header = T, stringsAsFactors = F, sep = ";")
defaultData <- subset(defaultData, select = c("CVR", "default", "Normal_slut", "Normal_start", "Type"))

defaultData[,c("Normal_start", "Normal_slut")] <- lapply(defaultData[,c("Normal_start", "Normal_slut")], funct

#Merge
totalSet <- merge(x =
totalSet$CompanyAge <- as.numeric(difftime(totalSet$GivenEnd,totalSet$Normal_start,units = "days"))/365
totalSet$ExecutiveAge <- totalSet$age

```

2. Data

To predict the future state of companies we strive to construct a rich data set, which both contains new and conventional data. It is essential that we observe characteristics of both surviving and bankrupt companies in a consistent manner to avoid any biases[1]. To avoid any data specific biases, we extracted data from the Danish Business Authority and several other data sources. This process of extracting and merging is outlined in the flowchart in Figure 2.1 below.

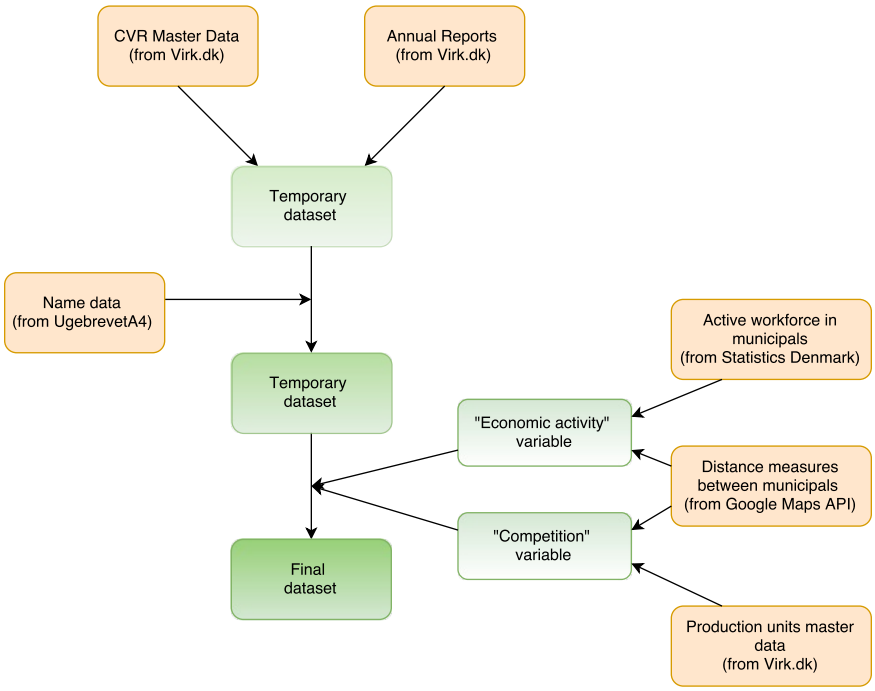


Figure 2.1: Flowchart showing how the various data sources are merged

2.1 Extracted data from virk.dk

Extracting data from the Danish Business Authority (DBA) is not trivial. In order to get CVR master data and annual reports on Danish companies, we created an account at the DBA. This granted us access to their API's. We created several solutions in Python that could extract the desired data from DBA by sending specific requests based on ElasticSearch in a JSON-format to the API's. We extracted desired data on all Danish companies. We also included annual reports from xml-paths using the *ÅRL taxonomy* from DBA and information on which companies were started, active or bankrupt in the period 2013-16. The Python solutions of this comprehensive task can be found in the separate code appendix.

[1]The data sets provided for the case, did not meet these criteria. The CVR master data did not contain defaulted companies, which of course imposes a serious bias. The FSA data set seemed inconsistent, which also posed serious problems for modelling. However the data we extracted is of the exact same nature as the platform data set "Annual reports" and can therefore be seen as the same source. For a more detailed overview of data problems see Appendix A.

2.2 Danish Business Authority: Master data and annual reports

By using the offered API's from the DBA as described above, we were able to extract CVR, P units, sub-industry, municipals, start date, end date and default for each company. We defined the status 'default' as a company experiencing either bankruptcy or forced liquidation, since the latter often leads to bankruptcy later on and in any case can only be undesirable for a bank. Furthermore, we extracted $\sim$ 630.000 annual reports for the years 2013-2016.

2.2.1 Financial ratios

We used the DBA data to create the financial ratios Return on Investments, Liquidity, Return on Assets and Gearing, because including these ratios could give a predictive performance of the models not obtainable by the variables separately. One of the main reasons for choosing exactly these ratios were that the inputting variables had very low percentages of missing data (see Appendix A.4). Furthermore these were the ratios found significant for Danish companies by Jensen (2013). The choices were also motivated by general findings in academia[2].

2.2.2 Changes in leadership

We created a variable that indicate whether or not the chairman of executive board had changed since the previous year. A change of chairman might be an indicator of bad management, dissatisfaction between chairmen and stack holders, or insider knowledge of distress.

2.2.3 Changes in audit firm

Similarly we created an indicator of change in audit firm. Assuming that the previous audit firm had annotations for the annual report this could be signs of potential problems in the future or other compliance indicating distress or bad management.

2.3 Name data

We also merged aggregated data on the first name belonging to the directors on the DBA data set. This data came from Ugebrevet A4, that had bought the detailed name data from Statistics Denmark. The data set contained data on *average* age, income, conviction etc. for every common first name in Denmark.

We included the name data because of the assumption, that the executive board often is a significant part of the reason for the default of a company, and that personal information on the chairman therefore could be an important predictor of a bankruptcy. A strong improvement would of course be personal information on these individuals, but this was not possible, therefore we use this data as a proxy.

[2]Beaver (1966) showed, that the development in financial ratios are different for companies up to default compared to non-defaulting companies, and Jensen (2013) showed, that this is still the case in Denmark today. Beaver et. al. (2015) found, that most studies have roughly equally good prediction power regardless of what ratios are used. Lundqvist and Strand (2013) later showed, that the financial ratios' influence on defaults have varied over time.

2.4 Spatial economic activity

We constructed a variable that accounts for spatial economic activity. The variable describes how dense the economic activity in each municipality is, taking spatial proximity into account. Economic activity is measured by local active workforce[3]. First, we determined how large the workforce is per radius in kilometers in each municipality. Secondly, we allowed for proximity effects between all municipalities and measured how the workforce per distance in kilometers from one municipality to all other 97 municipalities. To calculate the distances we used the Google Maps API. By sending municipality names, the API returns coordinates, which we used to calculate distances with the Haversine-Formula.

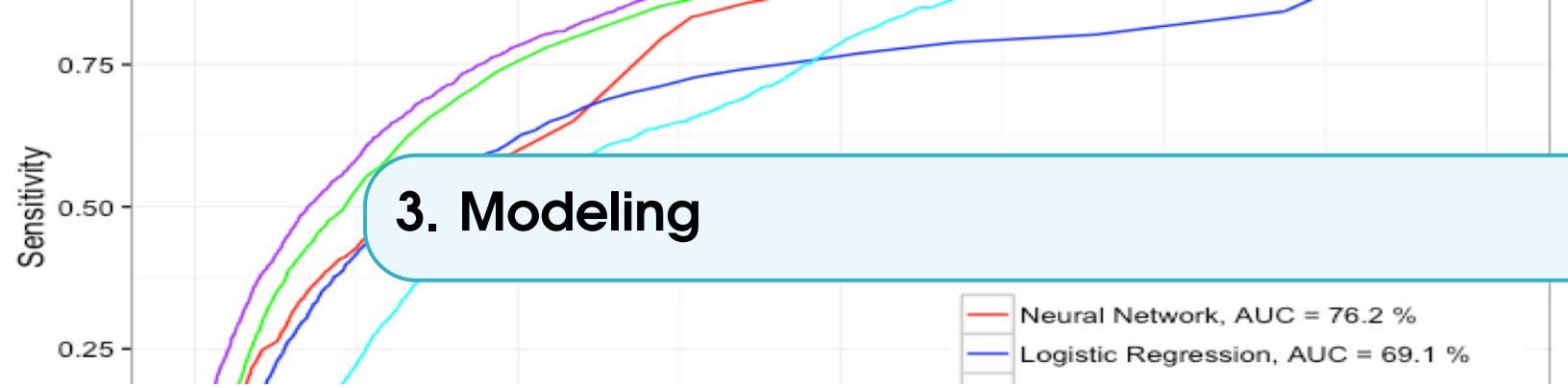
The economic activity variable was created for several reasons. Default rate could change because of different regional trends. Furthermore certain sub-industries could flourish in areas with high economic activity, while other sub-industries would flourish in the outskirts.

2.5 Competition

Using the CVR master data, we looked at how many local competitors the companies faced. For each municipality we counted the number of registered companies within the same sub-industry. Using the radius of the municipalities, we then calculated how many competitors each company face per radius km.

The variable was created to show the structural influence of the competition in different industries. Competition effects might both be positive and negative. Positive synergy effects and a high competition could be a result of a high demand, but on the other hand could hard competition shut down inefficient companies.

[3] Statistics Denmark RAS301



3. Modeling

In the following, we present the overall modeling strategy together with concrete models and our feature engineering process used to build and evaluate predictive models for company defaults. A flowchart describing the data analysis and model building process, can be seen in Figure 3.1.

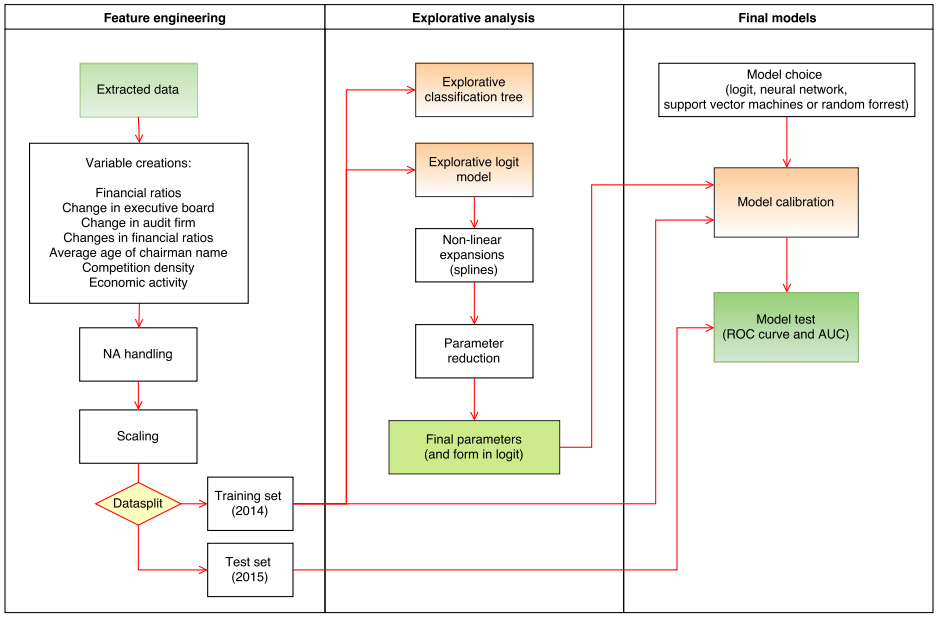


Figure 3.1: The flow from the extracted dataset to the final models

3.1 Modeling strategy and evaluation

For a model on bankruptcy to be useful to predict future defaults, the model has to be consistent over time. In order to predict anything useful in a given period, one has to use a model trained and fitted solely on a previous period. All models in this analysis are trained on annual reports where the report year ended in 2014 and tested on reports with an end date of the report year in 2015, since these are the latest years that can be trained and tested on.

The default variable deserves a few extra words, as this is obviously paramount to the model. In our calibration, a given annual report is flagged for default, if the company experiences a default within the 5-17 months following the end of the annual report. A more proper name for the default variable would have been "default within 12 months after publication of the annual reports". The 5 - 17 months stems from 1-12 months (one year) plus a 5 month lag: Companies in Denmark have to submit their annual report to the Danish Business Authority within 5 months after the end of the reporting year, so we can only assume to have information available after 5 months.

Although this modelling decision only makes the predicting harder (given an annual report, the predictive power will typically decrease over time), it seems to be the only reasonable thing to

do in order for the model to be of practical use. Consider for an example a company which by 31.05.14 delivers an annual report for the period 01.01.13-31.12.13. Once the report is made public on 31.05.14, there is no sense in predicting defaults for the period 01.01.14-31.12.14, for this is already half way over. Instead, once the '13 figures are in, the model predicts a probability of default within 01.06.14 - 31.05.15.

We want to emphasize that all sorts of model selection procedures and model training are done solely on training data, and the model performance reported are obtained by testing our trained models on a test data set that has not in any way been involved in the tuning or training of the models. It would be possible to achieve much higher predictive performance, if one allowed for model tuning and calibration on the test data. However, this would only be a result of over fitting to the present data, and would not in practice give a good predictive model. Similarly it would be possible to achieve higher predictive performance, if the models were trained and tested on data within same year.

Not only do we perform the model selection, tuning and training in a disciplined way, we also make a point in mimicking the real world situation, and this makes the analysis more relevant and applicable.

3.2 Feature engineering

Since some of the variables were very heavily tailed, we applied a scaling to selected variables for the relevant models in order to improve predictive performance and training time. Depending on the range of the variables, we applied either log-transformations or root-transformations.

Another issue concerning data, was how to handle missing values. Some variables are most likely not missing at random, since one could suspect that companies with certain variables indicating distress, are more likely not to report these if possible. For that reason we developed a process to include missing values as predictors in the models. For non-numeric variables we included missing values as a label in the variable. For each numerical variable, we constructed a 0-1 variable coding missingness in that variable. By making interactions, we can effectively model missingness as a predictor.[1]

Furthermore, the variables mentioned in Chapter 2, e.g. financial ratios, spatial economic activity and name data, were also a part of the feature engineering process.

3.3 Exploratory models

In the following section we describe two exploratory models applied to give a descriptive analysis of the data and to screen for strong predictors to use in the subsequent predictive machine learning models.

In order to screen for strong explanatory variables, we first fitted a classification tree to our data. A classification tree produces a structured flowchart giving a visual representation of the importance of the variables. That way we got a better understanding of the hierarchy of the individual variables. This also gave intuition behind the mechanism of defaults, which is useful in understanding, interpreting and building more complication prediction models.

With the classification tree, we also examined, whether it makes sense only to include the company type as an additive variable, or if one should use completely different models for different company

[1] For concrete implementation strategies the reader is referred to the attached code.

types. One could suspect, that completely different dynamics determine if a pizzeria, a bank or a fishery goes bankrupt. To test if this is the case, we calibrated a classification tree to the training set for different industry types.

We also did a classical logistic regression analysis to get an understanding of the significance of the effect of each of the variables on the default. With a logistic regression we could examine non-linearities and the significance of the predictive variables. The significance was examined with confidence intervals. We decided to include non-linear effects of certain variables using a natural cubic B-spline basis. This was done in order to make the models more flexible to complicated variable-response relationships, and whether or not to include non-linear effects was determined using diagnostics plots based on residual analysis. Besides using the logistic regression as an exploratory model, we also made use of its predictive capabilities. The other predictive models are described next.

3.4 Exploratory results

Below we report and explain our findings from the exploratory data analysis, and how these were used to guide further model building.

3.4.1 Classification tree

The results from the classification trees for two of the industries can be seen in appendix B. Generally, the classification trees show, that it is almost the same variables that determine default across industries, and therefore, the industry type was chosen to be included additively further on, though it might seem counter intuitive. As expected, the financial ratios are very important along with the key indicators ProfitLoss and Assets. Interestingly, the average age of the chairman of the executive board also seems important. A higher age seems to cause a smaller probability of default. The competition density is also included in the tree for *Hoteller og restauranter* but in the way such that a higher competition leads to a smaller probability of default. This indicates synergy effects and that the probability of default in general is lower in crowded regions due to a higher demand.

3.4.2 Logistic regression

The explorative logistic regressions showed a non-linearity in both gearing and the solidity ratio around zero (see appendix C). This indicates, that the models used should be different depending on whether the equity is positive or negative. This can be solved in two ways. Either by splitting the data set into positive and negative equity or by expanding the variables with natural cubic B-splines and thereby catching the non-linearity. In this analysis, the latter was done.

A confidence interval showed significance of the variables shown in Table 3.1. These variables seem somewhat similar to the most significant ones chosen by the classification tree, and all the variables have the expected influence on the probability of default. Furthermore we see, that the significant parameters found are not solely from financial statements. The logit model finds variable for economic activity significant (the one made from extraction from Google Maps API) along the created factor describing, if the company has changed audit firm or not. The exploraty logit model also finds the industry types *Bygge og anlæg*, *Handel*, *Hoteller og restauranter*, *Industry* and *Transport* to default above average and the industries *Vandforsyning og restauration* and *Finansiering og forsikring* to default less.

Table 3.1

Predictor variable	Influence of an increase in the variable on the probality of default
Equity	Negative
Profit	Negative
Assets	Negative
Economic activity	Negative
Company age	Negative
Gearing	Positive
Change in audit firm	Positive

The rest of the variables were found insignificant, but this was found to be because of collinearities inbetween the predictors. Therefore in order not to risk excluding important variables, we chose to keep them all.

3.5 Predictive models

Based on the findings in the preliminary analysis, we decided to build suitable prediction models using both methods from classical statistics and various machine learning techniques.

3.5.1 First layer models

Firstly, we calibrated a standard logistic model with the non-linear expansions from the exploratory analysis.

Secondly, we chose to train a random forest on the data. We use random forest to avoid the overfitting phenomenon that is usually connected with using single classification trees. Hyperparameters for the random forest, i.e. number of trees to grow and number of variables to try at each split, were determined by cross-validation on the training set. Subsequently we used these hyperparameters in training of our random forest. However, the training was not done on the entire training set. Due to imbalance in the default classes, this would lead to a biased classifier. Therefore, we chose to train on a data set containing all of the defaulted companies, and then resampled the same number of non-defaulted companies at random from the training set in order to get a representative data set.[2]

Thirdly, we decided to fit a neural network to the data. We used a simple architecture using a single hidden layer and softmax output for the final layer. The number of neurons in the single hidden layer, was determined by cross-validation on the training set. This simple architecture was chosen in order to make training of the network feasible and to avoid overfitting to data. Given more computational power it would be possible to experiment with deep architectures of the network and different activation functions between the hidden layers. This would undoubtedly give the network more flexibility, and could in practise increase predictive performance. The neural network was trained on the entire training set.

Fourthly, we also fitted a soft margin support vector machine to the data. We chose the kernel, kernel parameters and margin parameters by cross-validation. Like with the random forest, we also chose to do training on a data set with equal number of defaulted and non-defaulted cases.

[2]For reproducibility the all sampled data sets were generated using a seed.

Figure 3.2 shows the flow of the calibration of the three machine learning models used.

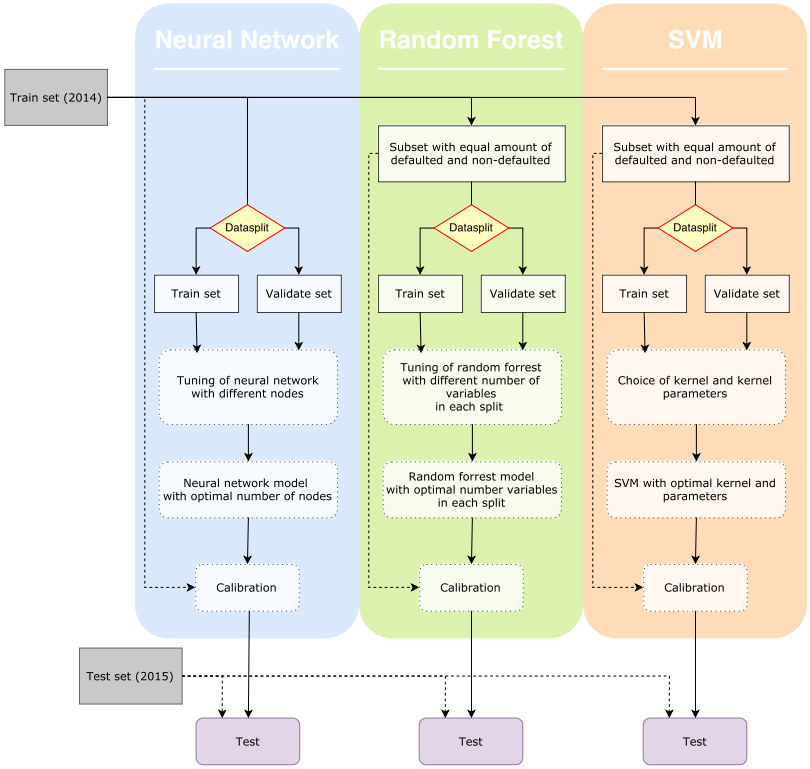


Figure 3.2: Flowchart showing the calibration and test of the machine learning based models

3.5.2 Ensemble Model

On top of this, we fitted an ensemble model. It can be shown that an ensemble of models will perform at least as well as the best first layer model in the ensemble, and therefore it seemed natural to combine our models in order to obtain a super-model, which utilizes the strengths of all other models.

As our ensemble method we chose to use stacking, inspired by its impressive performances in machine learning competitions such as the famous Netflix movie ranking competition. Instead of naively doing a basic model averaging with some predefined weightings of each model, the method of stacking is a model based approach, which yields the optimal weighting of each model in the ensemble. In stacking, the output of the individual models are fed into a combiner model, which then acts as the final prediction model. In our case, each of the models produces a probability of default for each of the companies. For the neural network we use the softmax output as probabilities, for random forest we use vote proportions, and for the support vector machine we use Platt scaling. We then used a logistic regression as our combiner model, which has as input the probabilities from the other models, and then finds the optimal weighting of these probabilities, to produce a final probability of default. The flow of the calibration and test of the ensemble model can be seen in Figure 3.3.

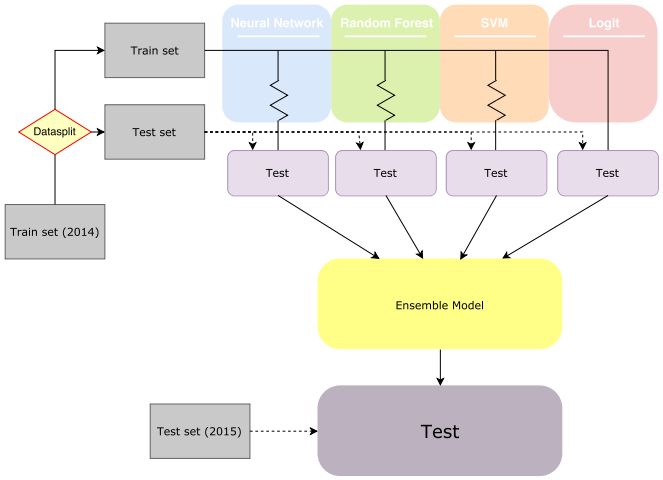


Figure 3.3: Flowchart showing the calibration and test of the ensemble model

Apart from having the primary objective of producing better predictions, using the ensemble model brings a secondary benefit: For machine learning algorithms, the predicted probabilities are often just intensities on some scale which quantify the risk of defaulting, but if e.g. the neural network produces a value of 0.3 there is no theoretical evidence of a 30% probability of default - especially not, if the models are trained on sets with high concentration of defaulting companies. However when our terminal model is a logistic regression on a full set, the predicted probability can be understood as an actual probability of default. Not only can the model predict defaulting companies at some threshold on an artificial scale, with the strength of the ensemble model it will also produce actual probabilities of default.

3.6 Predictive results

We present the predictive performance using ROC curves, AUC values and sensitivity (true positive rate) given a required level of specificity (true negative rate).

The logistic model was calibrated to the variables with the non-linear extension chosen in the last section. For the random forest, we choose to grow 1000 trees and randomly sample 4 variables in each split. The neural network performed optimally with 8 neurons in the hidden layer, and the soft margin support vector machine performed best with a linear kernel and a cost parameter of 1. The ROC curves from the model test of the four first layer models and the ensemble model can bee seen in Figure 3.4.

We see that the best performing models are the ensemble model and the random forest model. They have the highest sensitivity across all levels of the specificity and the highest AUC value. They predict almost equally good, which comes from the fact, that the random forest predicts better than the rest of the first layer models across all levels of specificity, and therefore the predictions of the ensemble model is based almost solely on the predictions from the random forest. However

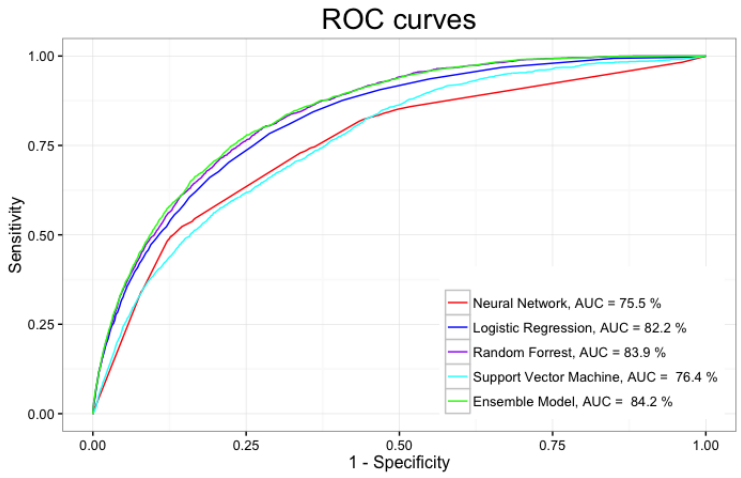


Figure 3.4: ROC Curves together with AUC values for each of the five models trained on 2014 and tested on 2015.

the ensemble model has the huge advantage over the random forest that it is based on a logistic regression on the full data set. This entails, that the probabilities given by this model are very reliable, whereas the probabilities given by the random forest are far too high due to the fact, that this was trained on a smaller and more balanced set.

	Specificity	60%	70%	80%	90%	95%	99%	AUC
1	Logistic Regression	87.1%	79.3%	67.1%	48.3%	33.0%	11.8%	82.2%
2	Neural Network	78.4%	68.8%	58.1%	41.0%	21.7%	4.5%	75.5%
3	Support Vector Machine	77.3%	67.4%	56.3%	38.9%	24.6%	5.4%	76.4%
4	Random Forest	88.8%	81.4%	69.9%	49.9%	35.0%	11.7%	83.9%
5	Ensemble Model	88.9%	81.8%	71.2%	51.8%	35.5%	12.4%	84.2%

Table 3.2: Sensitivity for each of the models given a specific level of specificity.

To illustrate the predictive power of the models, Table 3.2 shows the sensitivity of each model at six different levels of specificity. This shows, that if we allow a specificity of 70% i.e. we only allow for a 30% false positive rate, we are able to predict more than 80% of the defaulted companies with the ensemble model. If we only tolerate 5% false positive rate, we are still able to detect incredible 35.5% of the defaulted companies.

3.6.1 Over-time-consistency: Applying to 2013-2014

The model is only usable, if it is time consistent. To validate if this is the case, or we were only lucky between 2014 and 2015, we ran the full framework again, this time using 2013 as training year and 2014 as testing year. If the constructed modelling algorithm only describes specifically 2014-2015, the framework should have performed badly on 2013-2014. However, this is clearly not the case. On a test on 2014 data based on training on 2013 produces AUC's for the random forest and ensemble models at 84,6% and 82,2% (see the ROC curves in Appendix A.6). Both of these thus prove to be consistently high over time, which serves as a great validation that the model is managing to describe the general pass from one year to the other.

4. Business value and applications

In the sections below we discuss what business value Danske Bank can draw from the more academic explorations in the previous sections. The following two sections discuss what direct value to Danske Bank the data and modelling work can have, whereas the third section examines ideas that could be the base of further applications.

We develop the Probabilistic Profit Harvesting Tool (PPHT), which based on financial figures develop risk-weights and measure them against the ensemble model predictions. We show that Danske Bank can increase their profit on business loans with 1.3 %.

We suggest applications that could improve Danske Bank's market position as a leading technology based bank. For instance a customer benchmarking tool, which helps Danske Bank to easily deliver market analytics to all Danske Bank's business clients. Another suggestion is the easiest default probability reduction for a data-driven client advising tool, suggesting concrete improvements, lowering customer defaulting probability, and benefitting both customers and Danske Bank.

4.1 Value of the dataset

One major advantage is our holistic data approach. The process provides a fully automated baseline solution with strong predictions based on only publicly available data. The data extraction process we have developed is a free and fully automated way of producing a vastly strong data set. As the code can run from scratch without any manual labour, it is easily updatable when new annual reports are published or companies go bankrupt. We consider the data set a significant part of our product, because it allows for several valuable side applications.

For one instance, it will provide a valuable source for Danske Bank in doing market research in many areas besides modeling economic performance on companies. This could be anything ranging from statistics on precise geographic location of every company in Denmark to every audit firm's client database. The data also contains the bank of every company, so it would also be possible to do statistics on, e.g., all Nordea's business costumers even within every industry. There are comprehensive opportunities in the data set that contains more than 34 million non-Na data points, which can be of most significant value both to Danske Bank, KMD, Microsoft and the society through the knowledge that can be obtained through this data set.

4.2 Value of the models

4.2.1 Quantification of credit risk and data improvement potential

The most obvious value to Danske Bank is the quantification of their credit risk. It can help Danske Bank to quantify the probability of default for individual customers that are requested in the Basel Accords. Our final ensemble model has huge default prediction power, which is much better than those obtainable with the logistic regression, the neural network or the support vector machine. If Danske Bank has not already implemented an ensemble model on top of a random forest or more

advanced models, it could be of great value to add these methods to their default assessment tools. This could help with more precise credit rating of their business customers and better assessment in the decision making when companies apply for credit.

The model is very applicable in practice, as it pays attention to the time perspective, thus producing a framework for predicting a default 5-17 months after the end of a business year (i.e. from publication of the annual results and one year ahead). Furthermore the model is very consistent over time, such that it can be used to predict future defaults based on training on previous data.

4.2.2 Improvement potential

We imagine that Danske Bank within the same conceptual framework, but using their higher degree of data quality in their own customer base, could implement more sophisticated quantitative variables. This would e.g. be in the lines of the data set 'Behaviour' for old costumers, which we unfortunately were not able to include, since the set was for obvious reasons was anonymized.

We also highly recommend Danske Bank to invest a considerable amount of energy in getting detailed data on the board of directors. When the classification tree can find something as vague as the average age of the name of the chairman in the executive board important, more detailed data will definitely improve the bankruptcy model. One could imagine that data such as actual age, professional network, income, education etc. of all board members would have significant influence.

Since the model is extremely time consistent, it might also be able to predict defaults even further away in the future, such that the bank has longer time to take the required actions to minimize losses or maybe even prevent the default.

4.3 Value applications

In this section we suggest numerous ways to apply the value gained from the bankruptcy ensemble model in practice and thereby add significant value both to Danske Bank but also to its business costumers.

4.3.1 The Probabilistic Profit Harvesting Tool

The ensemble model outputs a probability of a company defaulting within year. However, a probability of default cannot alone determine whether a company should be granted (or refinanced) a loan or not.

It is crucial to evaluate the economic profit of the model. Whether a loan should be granted or not depends on the potential costs associated with each of the decisions. These are two fold: i) if a loan is given and the company goes bankrupt, the creditor suffers a loss. ii) if a loan is not given to a surviving company, the creditor suffers an opportunity cost. So the optimal exclusion rate on a loan depends on the cost of making a Type 1 error (loaning to a defaulting company) and the cost of making of Type 2 error (not loaning to a non-defaulting company). The optimal exclusion (the specificity) rate is thus company specific, and it is not easy to determine this value. To tackle this problem, going from a given probability to a decision on whether to grant a loan or not, we have developed a model we call the Probabilistic Profit Harvesting Tool or PPHT model[1].

[1] One could note the striking resemblence with the last names Pedersen, Petersen, Harmat and Thams.

The PPHT model will give a precise profit improvement by using our ensemble model, benchmarked up against refinancing or granting a loan to every company. Notice that the process of using our ensemble model combined with the PPHT model would not require any human interaction, i.e. a local bank advisor's assessment of financial statements and etc. The model is fully described in Appendix A.5.

In the model, we assume, the following:

- Danske Bank is risk averse.
- Old and new potential customers are representative of the population.
- The company only has one creditor (Danske Bank).
- Every loan has to be refinanced once in a year.
- If Danske Bank refinances a loan and the company goes bankrupt, they can recover up to 60% of the assets.
- If Danske Bank refinances a loan, and the company does not default, they earn 7% in interest.
- If Danske Bank refuses to refinance a loan, they lose money equivalent to 0,01% of the assets in the company on *bad publicity* or *reputation cost*. This is apart from the opportunity cost.

With these assumptions, we calculated the optimal credit decision for all companies. The calculations lead to the the result, that Danske Bank should refuse to refinance or give loans to 8% of their costumers with specific characteristics. Among these 8% refused companies, Danske Bank would predict 28% of the defaulting companies correctly and not suffer any loss. While accepting all costumers Danske Bank would suffer a loss in 1% of cases, but they would earn interest on all the remaining.

By implementing the models Danske Bank would increase their profit with 1.3% relative to the benchmark, while lowering their risk considerably. In other words the reward of the ability to predict some defaulting companies is larger than the loss of refusing surviving ones.

4.3.2 Easiest default probability reduction

In the current framework of predicting probabilities we propose another tool for Danske Bank to explore: The easiest default probability reduction.

Imagine a situation where Danske Bank has trained a framework for predicting default probabilities, and are considering how to bring down the credit risk exposure for a specific client, say company X. Within the trained model, company X has an assigned probability of default, p_D. Danske Bank could now use the trained model, to make numerical estimations[2] of which of the input variables would be most beneficial to change in order to bring the model probability prediction down. Assuming a causal relationship between specific financial ratios and the probability of default, this would give Danske Bank a competent and clear guidance-tool for advising their clients, benefitting both customers and Danske Bank.

We believe it can be a valuable insight to Danske Bank to be able to say to company X: "The model predicts a probability of p_D and the model believes that the easiest way to reduce this risk is to increase (e.g.) Current Assets". This would be an easily applicable tool in helping customers understand to which financial figures they need to spend attention and can help to bring down the overall default rate among Danske Bank customers.

[2] In a very practical implementation one could create data lines where each line had just one variable scaled down or up, and then use the model to predict on these data lines, to see which change produces the largest decrease in probability.

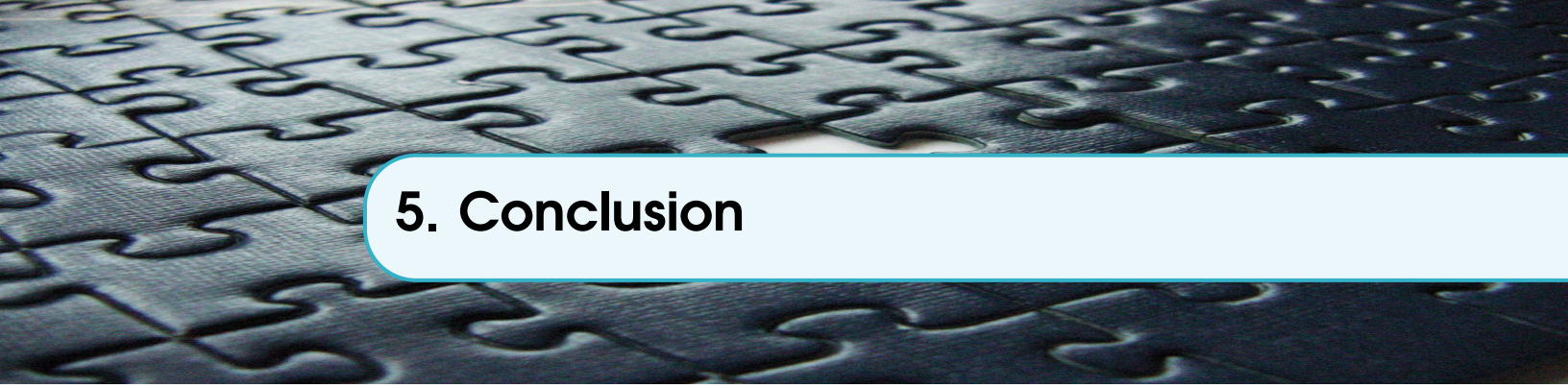
If Danske Bank wanted to build this framework even further, using methods from causal inference, it would be possible to build models for predicting effects of interventions on specific financial ratios on the probability of default.

4.3.3 Customer benchmarking tool

One way the data set can create impact to individual companies would be to create a benchmarking tool. We imagine a small carpentry company, where the owner would like to see how his business is doing compared to the company's competitors. Danske Bank could easily extract economic figures for companies in the same industry, of comparable size (economic or in number of employees) and in the same region (e.g. by zip-code). The carpenter could then see the company relative position in e.g. profit or return on equity and even compare to individual named competitors as well as being given the variables for economic activity and competition variables for his vicinity. All this information is already in the data set and would easily be aggregated. Wrapped in a pretty HTML-environment, this would provide a most insightful gift to the Danske Bank business customers with a minimal cost and effort to Danske Bank.

4.3.4 Other applications within Danske Banks business model

As the model is based on publicly available data, the model provides a method of assessing risk pools, which are not currently on Danske Bank's book, but that Danske Bank may wish to evaluate. This could either be with the prospect of buying up portfolios of loans, buying or selling structured derivatives or that Danske Bank would wish to asses the risk and estimate the value of a competitors book to know the competitors position. Our approach would provide a strong tool in making Danske Bank take data driven decisions in such situations.



5. Conclusion

This paper delivers two key technical achievements:

Data A stable and fully automated framework has been set up, which extracts publicly available data and adds value by merging with several other sources. The data contains 34 million non-NA data points and we believe it can be the base of vast analysis, way past the scope of this paper and several analyses could be derived hereof. Within the scope of predicting default we suggest that investing in information on board members could be valuable.

Model The final bankruptcy model developed showed very high predictive performance and firm stability. This model was an ensemble model averaging a logistic regression, a neural network, a support vector machine and a random forest. Each of these were calibrated through a train and test procedure on different years. In an AUC test, the ensemble model obtained an AUC of 84,2% from calibrating on 2014 and testing on 2015.

In the process of developing the model, we have learned quite a few methodical elements, which we believe is also of value to Danske Bank. Among others, we provide a method to pull data from virk.dk as described in Section 2.2, a method to include spatial economic activity as described in Section 2.4, show the fact that it is worthwhile to consider changes in audit firm as seen Section 3.4.2, develop a way of handling the NA-data as seen in Section 3.2, and show how an ensemble model on top of a random forest is the best default prediction model, that also produces correct probabilities. We believe that all this will provide model engineers at Danske Bank with plenty of inspiration to implement them as bricks in bankruptcy models and other constructions.

Considering the strict methodical discipline exercised such as over-year test/training split, NA-inclusion, unbiased model calibration and the 5-17 months default window used, the resulting ensemble model stands out as an impressively well performing model. Quite significantly this strictness pays off by delivering a framework which is over-time-consistent, mimicking the practical timing-problems of annual report publishing faced by Danske Bank and is thus easily implementable for Danske Bank.

The reliable and unbiased model would mean that Danske Bank could readily implement the framework or elements of it, and as discussed in Section 4.2.1, the model could become even stronger if non-public data were added to the data set. A number of ideas are also presented for how Danske Bank can take further advantage of the model, beyond the scope of quantifying credit risk. Among other suggestions are taking advantage of the predicting model to find which of the key variables would most effectively bring down the model probability of defaulting; and using the acquired public data set to provide a tool for benchmarking with a clients competitors.

The Probabilistic Profit Harvest Tool developed showed that Danske Bank could increase their profit on business loans with 1.3 % by refusing a refinance to 8 % of its clients. This could be done by applying an automated risk-weighted process for giving loans with minimal human interaction, freeing up employee time for other important tasks.

Bibliography

[1] David Lundqvist and Jacob Strand (2013), *"Bankruptcy Prediction with Financial Ratios - Examining Differences across Industries and Time"*, Master Thesis, Lund University

[2] William Beaver (1966), *"Financial Ratios As Predictors of Failure"*, Journal, Jornal of Accounting Research

[3] Peter Juel Jensen (2013), *"Konkursforudsigelse af danske virksomheder"*, Master Thesis, Copenhagen University

[4] William Beaver et. al. (2005), *"Have Financial Statements Become Less Informative? Evidence from the Ability of Financial Ratios to Predict Bankruptcy"*, Journal, Review of Accounting Studies

A. Appendix

A.1 Comments on the distributed data

As mentioned within the paper, the skeleton in the data set is the Annual Reports published by the Danish Business Authority. Our extraction of this is done in the lines of the set "Annual Reports" on the BAC platform, which contain cvr-figures and xml-links (Though we actually pulled out a longer list of annual reports, as described below). Apart from this all our other data sources we have pulled in ourselves, and in the following we will briefly describe the reasons for why it wasn't possible to include other of the platform data sets.

Danske Bank data

Two major unfortunates prevented us from applying the FSA data:

Anonymization The anonymization made it impossible to merge this data with any other public data. This made it impossible to add new data and original ideas to the models on the data sets, thus one could only hope to replicate models, that Danske Bank has already made. Also an important part of the development process was to spend (frankly quite a lot of) time digging down in the data to get an intuitive feeling of the data. Taking an sample company from the set, do the figures look reasonable? That question is much easier to answer, when you know whether you're looking at 'Hollywood Pizza' or 'Nordisk Fjer'.

Simulated noise Looking at the simulated data in the FSA, we saw several things that were a bit unexplainable to us. The distributions of the variables seemed far to smooth, and many variables made little economic sense. An example is seen in Figure A.1, which shows a density plot of the *5%* to *95%* quantiles on Current Assets in the data from Danske Bank and the data we extracted. From our understanding, Current Assets can very rarely (if ever) be negative, and this made us fear that the noise has pushed too many observations too far away from their starting point. Data accuracy in Current Assets is quite an important, since it is the numerator in the liquidity ratio - one of the most important ratios when predicting default.

In the view of our trouble with the FSA data, the Danske Bank behavioural data would have been standing a bit alone, and it is anyways still subject to the anonymization problem.

Public data

KMD had delivered master data on CVR numbers and P units but it is our impression that this data was only extracted for active firms, so unfortunately that made it unusable for our scope, modelling defaults.

The file "Annual reports" with xml-paths unfortunately seemed to have some flaws. The first occurence is in line 9474, where the pdf- and xml-path for *Bang & Olufsen* is missing. All xml- and pdf-paths are then pushed one line up, so the report in that line corresponds to the CVR number below - it is shown in the fig. A.2. This continues to happen several times through the file, such that the correct cvr and xml get further and further away from each other. The columns are still of equal

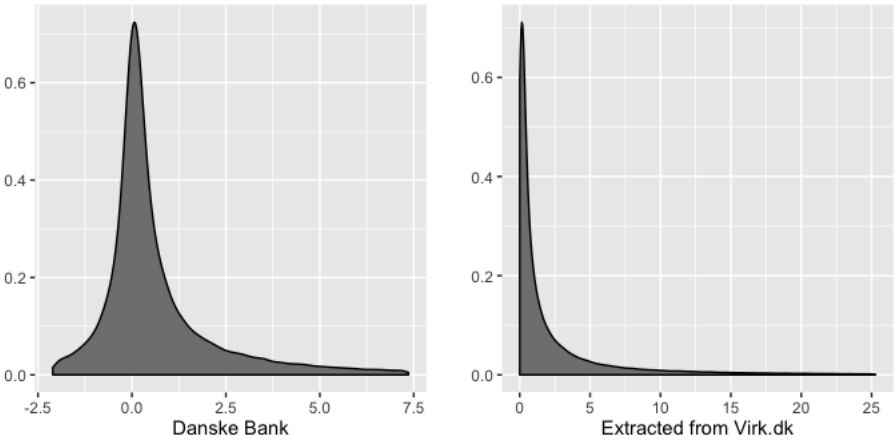


Figure A.1: Density plot for Current Assets in millions from both Danske Bank's data set and the extracted set from Virk.dk.

lengths so somewhere there must also be CVR numbers that are missing. Because of this, we found it easier to extract our own version of a similar file, than to be correcting the existing. That way we could also get twice as many reports, which were for some reason not in the platform "Annual reports" file.

	A	B	C	D
1	cvr_nummer	xml_links	pdf_links	
9472	34600791	http://regnsl	http://regnsl	
9473	32330487	http://regnsl	http://regnsl	
9474	41257911	http://regnsl	http://regnsl	
9475	15527137	http://regnsl	http://regnsl	
9476	15177284	http://regnsl	http://regnsl	
9477	28682506	http://regnsl	http://regnsl	

Figure A.2: First error occurrence in the annual reports file

The Septima data set seemed very interesting, but again there was only data on active firms and a few companies which defaulted within a very recent time horizon. This could therefore unfortunately not be used, as we needed to include all the defaulted companies for a larger time frame.

A.2 Classification trees

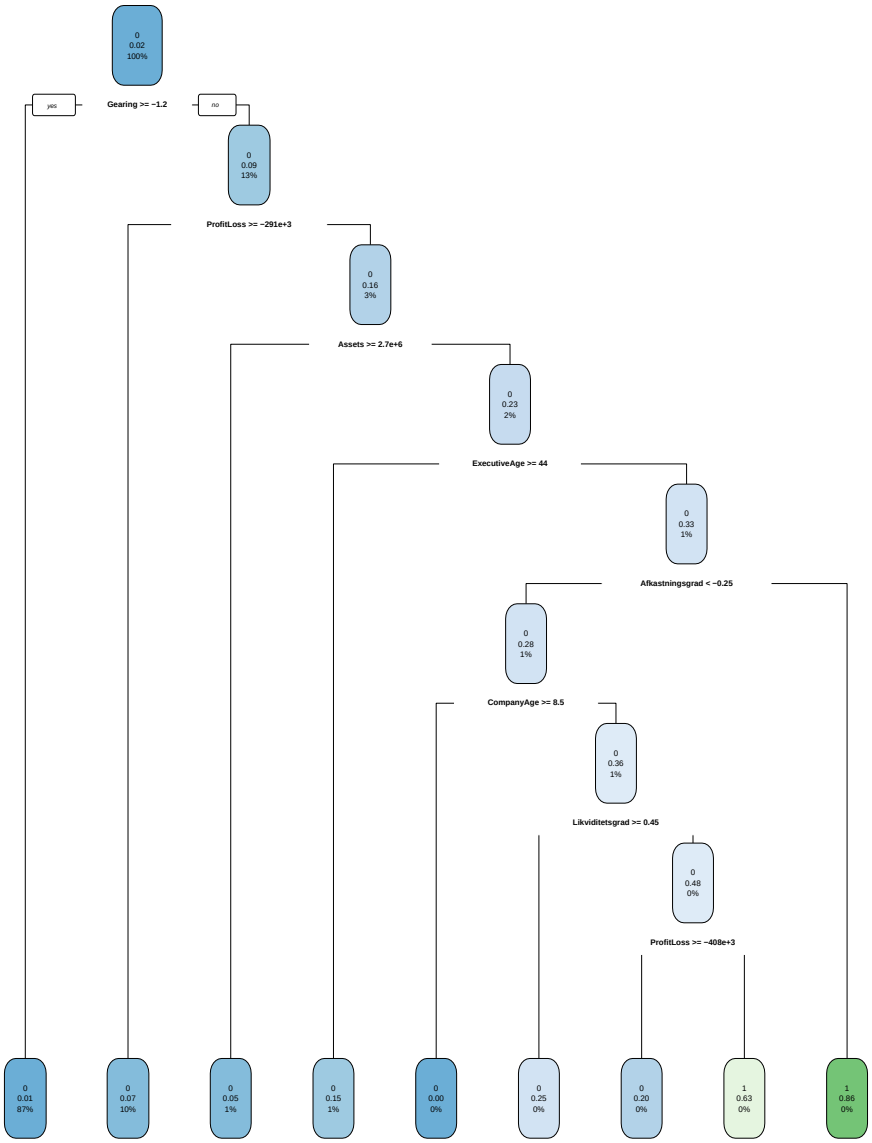


Figure A.3: The industry type *Bygge og anlæg*

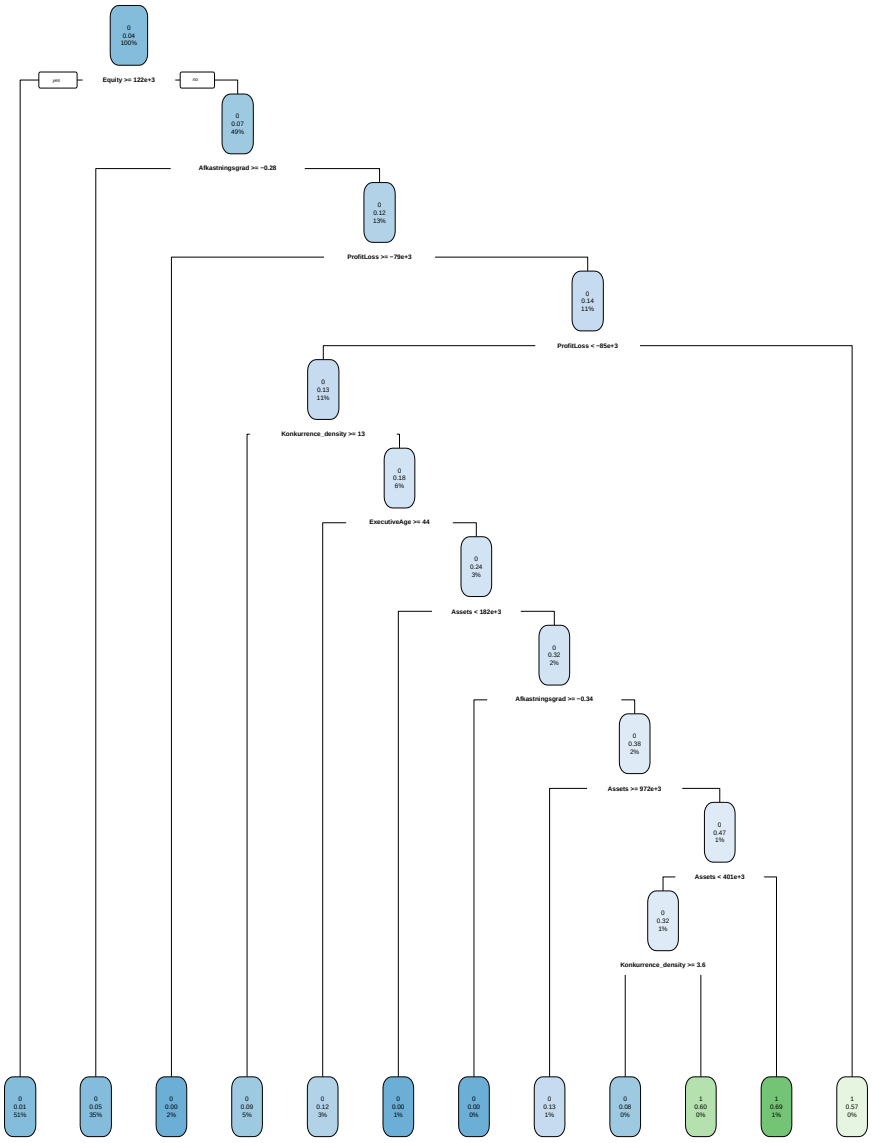


Figure A.4: The industry type *Hoteller og restauranter*

A.3 Explorative logit

The logit regression was first fitted with the following variable all entering additively:

- Equity
- Profit/loss
- Assets
- Cash and cash equivalents
- Industry type
- Company type
- Competition variable
- Economic activity
- Company age
- Average age on executive's first name
- Solidity ratio
- Liquidity ratio
- Gearing
- Change in solidity ratio
- Change in return on investment
- Change in liquidity ratio
- Change in gearing
- Change of audit firm
- Change of leader of executive board

When using logit model, we assume, that the conditional distribution of the response given the predictors is given by the exponential dispersion distribution belonging to the logit model.

An easy way to look for violence of this distribution is to investigate, if the residuals are dependent of the predictors. To examine this, we plotted the deviance against both the fitted values and each of the continuous predictors. The results for some of the variables can be seen in fig. A.5.

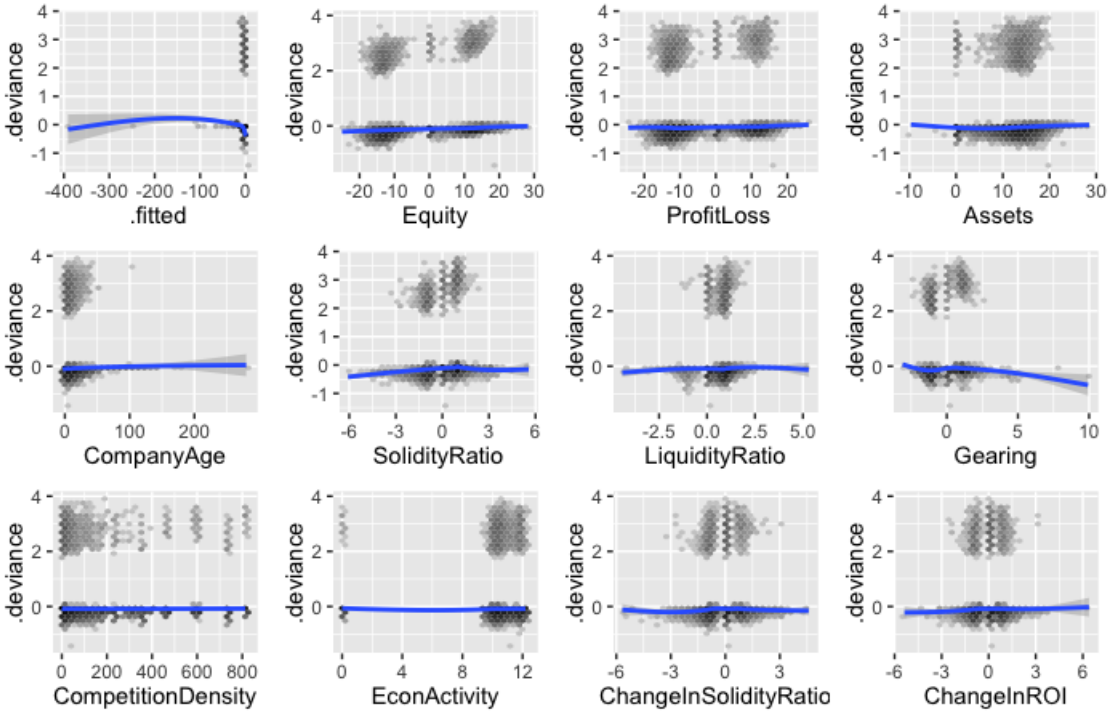
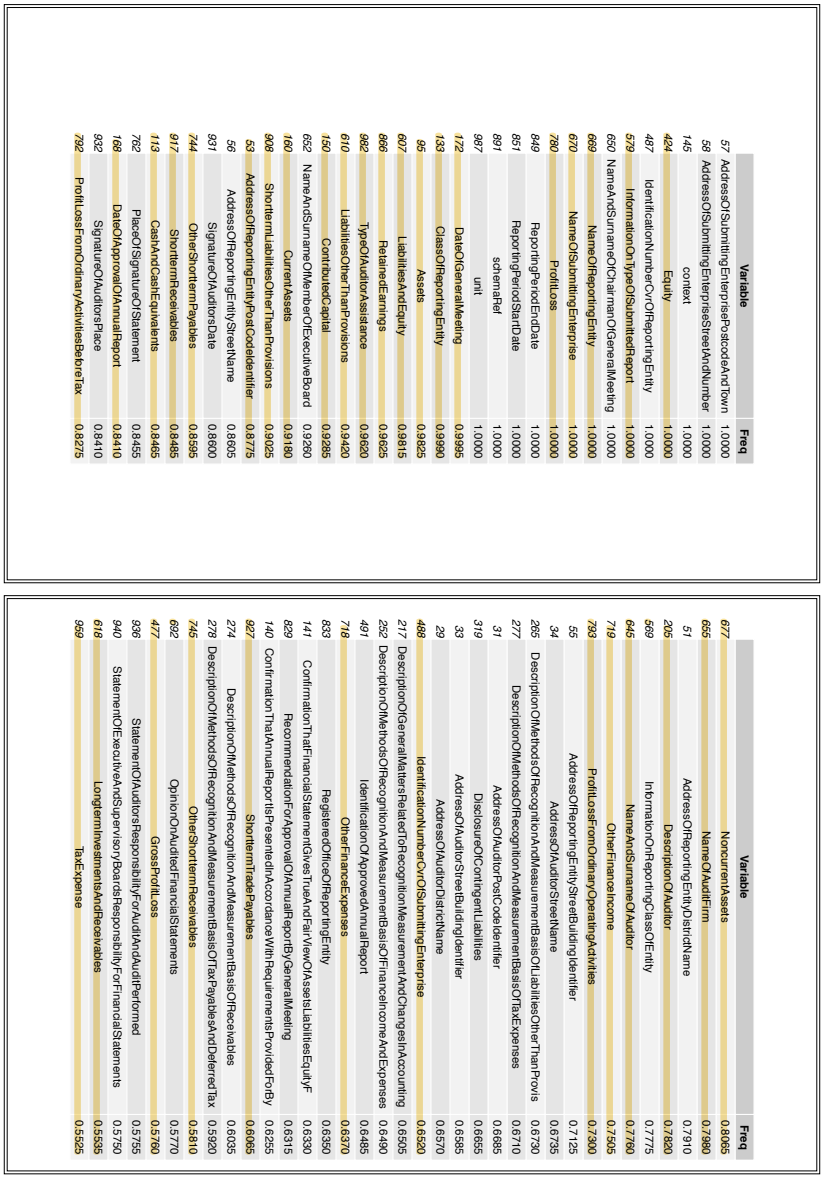


Figure A.5: First error occurrence in the annual reports file

As one can see, the smoothing line does not seem to be zero in all plots. In solidity ratio and gearing, we see a bend around zero. We solved this with a natural cubic splines with three degrees of freedom (that is, two internal knots). The knots were placed based on the quantiles to prevent overfitting the data.

The diagnostic plots afterwards were much nicer in solidity ratio and gearing, and the assumptions seemed to hold fairly well afterwards.

A.4 A selection of variables and their frequency



	Variable	Freq
57	AddressOfSubmittingEnterprisePostcodeAndTown	1.0000
58	AddressOfSubmittingEnterpriseStreetAndNumber	1.0000
145	context	1.0000
424	Equity	1.0000
487	IdentificationNumberCvrOfReportingEntity	1.0000
579	InformationOnTypeOfSubmittedReport	1.0000
650	NameAndSurnameOfChairmanOfGeneralMeeting	1.0000
669	NameOfReportingEntity	1.0000
670	NameOfSubmittingEnterprise	1.0000
780	ProfitLoss	1.0000
849	ReportingPeriodEndDate	1.0000
851	ReportingPeriodStartDate	1.0000
891	schemaRef	1.0000
987	unit	1.0000
172	DateOfGeneralMeeting	0.9995
133	ClassOfReportingEntity	0.9990
95	Assets	0.9825
607	LiabilitiesAndEquity	0.9815
866	RetainedEarnings	0.9625
982	TypeOfAuditorAssistance	0.9620
610	LiabilitiesOtherThanProvisions	0.9420
150	ContributedCapital	0.9285
652	NameAndSurnameOfMemberOfExecutiveBoard	0.9260
160	CurrentAssets	0.9180
908	ShorttermLiabilitiesOtherThanProvisions	0.9025
53	AddressOfReportingEntityPostCodeIdentifier	0.8775
56	AddressOfReportingEntityStreetName	0.8605
931	SignatureOfAuditorsDate	0.8600
744	OtherShorttermPayables	0.8595
917	ShorttermReceivables	0.8485
113	CashAndCashEquivalents	0.8465
762	PlaceOfSignatureOfStatement	0.8455
168	DateOfApprovalOfAnnualReport	0.8410
932	SignatureOfAuditorsPlace	0.8410
782	ProfitLossFromOrdinaryActivitiesBeforeTax	0.8275

	Variable	Freq
677	NoncurrentAssets	0.8065
655	NameOfAuditFirm	0.7980
51	AddressOfReportingEntityDistrictName	0.7910
205	DescriptionOfAuditor	0.7820
569	InformationOnReportingClassOfEntity	0.7775
645	NameAndSurnameOfAuditor	0.7760
719	OtherFinanceIncome	0.7505
783	ProfitLossFromOrdinaryOperatingActivities	0.7300
55	AddressOfReportingEntityStreetBuildingIdentifier	0.7125
34	AddressOfAuditorStreetName	0.6735
265	DescriptionOfMethodsOfRecognitionAndMeasurementBasisOfLiabilitiesOtherThanProvis	0.6730
277	DescriptionOfMethodsOfRecognitionAndMeasurementBasisOfTaxExpenses	0.6710
31	AddressOfAuditorPostCodeIdentifier	0.6685
319	DisclosureOfContingentLiabilities	0.6655
33	AddressOfAuditorStreetBuildingIdentifier	0.6585
29	AddressOfAuditorDistrictName	0.6570
488	IdentificationNumberCvrOfSubmittingEnterprise	0.6520
217	DescriptionOfGeneralMattersRelatedToRecognitionMeasurementAndChangesInAccounting	0.6505
252	DescriptionOfMethodsOfRecognitionAndMeasurementBasisOfFinanceIncomeAndExpenses	0.6490
491	IdentificationOfApprovedAnnualReport	0.6485
718	OtherFinanceExpenses	0.6370
833	RegisteredOfficeOfReportingEntity	0.6350
141	ConfirmationThatFinancialStatementGivesTrueAndFairViewOfAssetsLiabilitiesEquityF	0.6330
829	RecommendationForApprovalOfAnnualReportByGeneralMeeting	0.6315
140	ConfirmationThatAnnualReportIsPresentedInAccordanceWithRequirementsProvidedForBy	0.6255
927	ShorttermTradePayables	0.6065
274	DescriptionOfMethodsOfRecognitionAndMeasurementBasisOfReceivables	0.6035
278	DescriptionOfMethodsOfRecognitionAndMeasurementBasisOfTaxPayablesAndDeferredTax	0.5920
745	OtherShorttermReceivables	0.5810
692	OpinionOnAuditedFinancialStatements	0.5770
477	GrossProfitLoss	0.5760
936	StatementOfAuditorsResponsibilityForAuditAndAuditPerformed	0.5755
940	StatementOfExecutiveAndSupervisoryBoardsResponsibilityForFinancialStatements	0.5750
618	LongtermInvestmentsAndReceivables	0.5535
959	TaxExpense	0.5525

Figure A.6: Figures in yellow were not excluded at an early stage.

A.5 The PPHT model

The optimal specifity depends on the cost of making a Type 1 error (loaning to a defaulting company), the cost of making of Type 2 error (not loaning to a non-defaulting company), and the risk aversion. Below the cost of making a Type 1 error is denoted D. The assumed cost due to bad publicity of not refinancing a loan is denoted L, and the reward for correctly loaning to a non-defaulting company is denoted ND.

The model is to be seen mainly as a proof-of-concept of how to develop a model estimation tool which can make decision based on not only probability but also risk appetite in the model. However we are well aware that the concrete choices of functions D, ND, U, and L could probably be varied to better match the reality the bank is experiencing.

The model takes the following assumptions:

- The company only has loans at one creditor. - The bank estimates the loss given default, D. It is assumed that if a company goes bankrupt, the bank looses the debt, d, minus a fraction $rec\%$ of the assets, A. Let:

$$D := -(d - rec\%A)^+$$

- The bank estimates the benefit of giving a loan to a non-defaulting company, ND. It is assumed that if a company does not go bankrupt (within a year), the bank earns r in interest on the debt, d, (after administration costs). Let:

$$ND := rd$$

- There is a loss due to bad publicity whenever the bank decides not to refinance a loan. The cost from the bad publicity is assumed to proportional to the company assets (as a measure of its size) by a factor $-\rho$. Let:

$$L := -\rho \cdot A$$

- The bank is equipped with a HARA utility function given by

$$U(x) = \begin{cases} 1 - \exp(-x \cdot 10^{-6}) & \text{if } x \leq 0 \\ \ln(x \cdot 10^{-6} + 1) & \text{if } x > 0 \end{cases}$$

where x is the economic profit to the bank.

The bank wants to optimize the expected utility, so they should estimate the company as defaulting, if the expected utility of refinancing the loan is lower, than the expected utility of not refinancing (which is negative because of the bad publicity)

$$p_D \cdot U(D) + p_{ND} \cdot U(ND) \leq U(L)$$

The treshold probability, for a company is where the above holds with equality. Inserting the utility function and solving for p yields the threshold:

$$p_T := \frac{U(ND) - U(L)}{U(ND) - U(D)}$$

That is, for an individual company, we can use the developed models to compute the probability of defaulting and then apply a company specific cutoff, related to the company assets and debt.

If we assume $\rho = 0,01\%, rec\% = 60\%, r = 0,07$, we get the threshold:

$$p_T = \frac{1 - \exp(-0,01\% \cdot A) - \ln(7\% \cdot d + 1)}{(1 - \exp(-(d - 60\%A)^+)) - \ln(7\% \cdot d + 1)}$$

where d is the bank debt and A is the assets.

For any set of companies with corresponding probabilities of default given by our model, Danske Bank can now for each company compute the company specific probability threshold, and thus base its decision not on a global threshold but tailor the decision making process to suit a given risk aversion (here the HARA utility) and by varying the input functions D and ND, by e.g. factoring in more complex interest models or recovery rates.

An example of three loan decisions based on the PPHT model, and their economic consequence, can be seen in Table A.1.

Table A.1

Company	p_D	p_T	Decision	Actual event	Economic result to bank
A	30%	40%	Refinance	Default	D
B	80%	50%	Don't refinance	Default	$-\rho \cdot A$
C	5%	10%	Refinance	Non-default	ND

In a general form the economic result, R_i for a company i, to Danske Bank is:

$$R_i = \begin{cases} -\rho \cdot A_i & p_{T,i} < p_{D,i} \\ D_i & p_{T,i} \geq p_{D,i}, \text{Actual event}_i = \text{Default} \\ ND_i & p_{T,i} \geq p_{D,i}, \text{Actual event}_i = \text{Non-default} \end{cases}$$

A.6 ROC curves

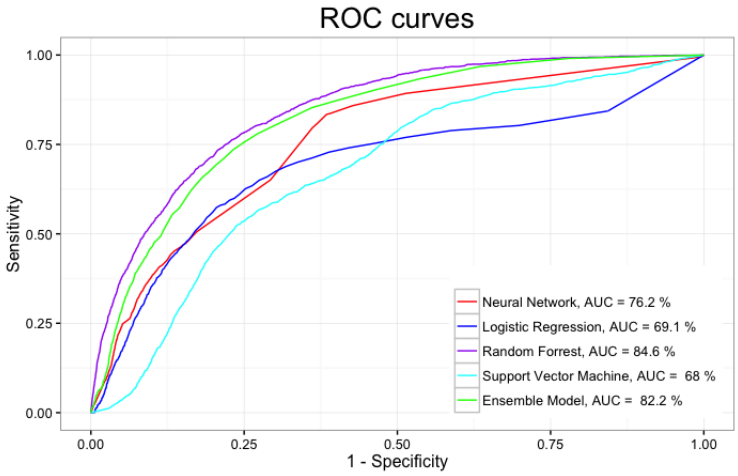


Figure A.7: ROC Curves together with AUC values for each of the five models trained on 2013 and tested on 2014.