AMSTERDAM
DATA
COLLECTIVE

DEVELOPING CREDIT SCORING MODELS

Title: An introduction to credit scoring models
Date: 2018
Author: Amsterdam Data Collective
F.A.O: New and potential consultants

This document is about scoring models; classification models that are used to calculate the probability of an event taking place and assigning a *score* accordingly. In other words, a scoring model is a model in which the dependent variable is categorical. Typical questions that would lead to a scoring model being developed are:

- Will prisoner X reoffend?
- Will patient Y's cancer be cured?
- Will I win my court case?

Each of these questions could be answered with a simple *yes* or *no* and ideally, we would provide such answers. But unlike Donald Trump, we are seldom able to provide answers with so much certainty. Instead, we provide probabilities that are based on explanatory factors. Finding out which factors contribute to the likelihood of the event taking place is what scorecards are all about.

Credit scoring models are a special case of scoring models and are aimed at answering the question: what is the probability of a client paying back his or her loan according to the agreed terms? Being able to answer this question allows our clients to manage their portfolios with confidence and reduces the risk of another financial crisis.

An introduction to credit scoring models

Credit scoring models are one of the "classical" yet still extensively used predictive modelling applications within finance. They are used by lending institutions to make an ex-ante decision on whether an extended credit will result in profit or loss. Though there are many forms of credit products, they all have in common that a lending institution provides a sum of money to a lender and expects this to be paid back at a later moment in time, including some amount of interest.

Before the lending institution takes the decision to extend a credit, it is essential to assess the likelihood of the lender paying back the loan and interest. This is done using predictive models. Thus, the goal of a credit scoring model is to discriminate between clients with different levels of credit risk. This is accomplished by estimating each client's *probability of default* based on a set of loan characteristics, such as time in arrears or loan purpose. How these loan characteristics are selected and related to the dependent variable is explained in more detail in the remainder of this document.

To develop a credit scoring model, data of historical lender behaviour is used. Such data sets contain information on repayment behaviour, in combination with a variety of explanatory variables, ranging from personal characteristics (e.g. age, income, education) to external variables (e.g. macroeconomic indicators). The dependent variable is binary: 1 if the lender defaults during the next period and 0 if he does not. This results in a model which can produce, based on information at time *t*, a prediction of the probability of default at time *t+1.*

Functional form

The credit scoring model assigns a credit rating to every client. This credit rating is directly related to a client's probability of default and is determined by estimating which client characteristics are more likely to lead to defaults. Because the target variable is binary, the relationship between the dependent and explanatory variables can be estimated with a binary regression such as the logistic regression:

$$PD = \Pr[default\ next\ period = 1|X] = F(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k), \qquad \textit{Equation 1}$$

where $default\ next\ period$ refers to the event in which a client goes into default during the next month after the snapshot date at which the dependent variable is evaluated, X to the set of loan and client characteristics at the snapshot date, β_0 to β_k to the coefficients to be estimated and $F(z)$ to the logistic cumulative distribution function (cdf), evaluated at $z = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$:

$$F(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}. \qquad \textit{Equation 2}$$

From these two equations it should be concluded that there are two challenges when estimating this relationship. The coefficients β_0 to β_k need to be estimated; however, to be

able to do that, the variables in X (x_1 to x_k) that best predict defaults need to be determined. The process of selecting the variables in X is referred to as the variable selection algorithm and is the topic of the next section.

Variable selection

The main challenge in developing a credit scoring model is to select those client characteristics that best predict whether a client will go into default. This could be achieved through a statistical variable selection algorithm, such as Stepwise or LASSO, but also by applying knowledge from experienced financial professionals. The disadvantage of simply using an algorithm, is that the variables might perform well within the obtained sample, but poorly when applied later on. The disadvantage of basing a model purely on expert judgement, is that experts might be biased or that their knowledge is outdated. This chapter outlines a process that combines statistics and expert judgement when developing a credit scoring model. This process is visualised in Figure 1.

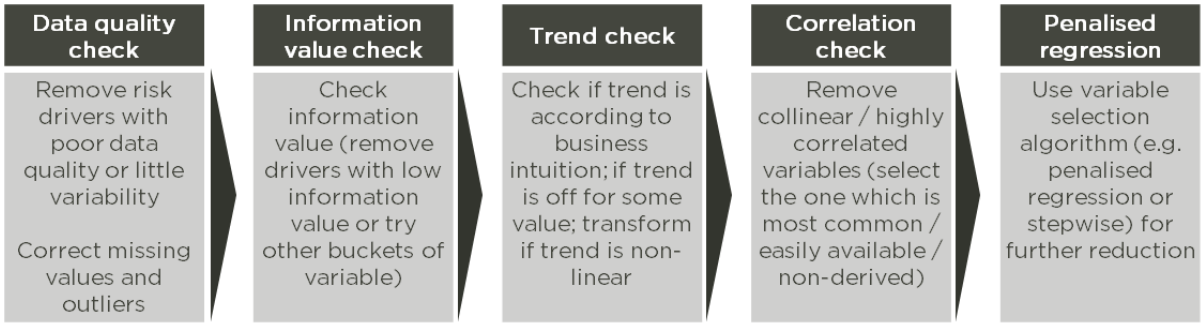

Figure 1: Visualisation of variable selection process.

By combining statistics with an expert-based approach, the probability of modelling the true relationship between defaults and client characteristics is optimised. Note that a plethora of alternative literature on estimating credit scoring models is available online.

Data Quality check

The first step of the variable selection procedure is to get rid of the variables with poor data quality or low variability. If data is not reliable, it cannot be used for modelling purposes, as this would lead to unreliable predictions. Before getting rid of unreliable observations, it is attempted to cleanse the data. There are many ways in which data can be cleansed, of which a few are mentioned below. When cleansing a data set, it is important to ensure that no bias is introduced and the maximum number of observations, and thus information, is kept.

Low variability

Variables with low variability, such as variables with the same value for every observation, are not useful as explanatory variables, as they cannot be correlated with the dependent variable, assuming that the dependent variable does fluctuate.

Missing Values

In practice, it will occur that variables contain missing values. A simple way of correcting these observations, is to remove the full record from the data set. This can, however, have significant impact on the final model outcome. It is therefore advisable to attempt to find out why a value is missing, before deleting the record. Alternatively, these variables could be recoded by creating an additional indicator variable or using the average or median as a proxy.

Outliers

Since the data sets used for credit scoring are a combination of manually entered and automatically generated information, they are prone to measurement error. For the detection of outliers, please refer to the vast number of methods available in statistical literature. Still, detecting outliers will always remain a subjective exercise; therefore, common sense is just as important and relying on statistical techniques.

Information value check

When deciding which explanatory variables to select, it helps to check each potential variable's information value. Variable with low predictive power should be disregarded. Variables can also be ranked based on their information value; variables with higher information value should be more likely to end up in the final model. The total information value of the categorical variable X is calculated as:

$$IV_X = \sum_{i=1}^{n} IV_{X_i}, \qquad \textit{Equation 3}$$

where IV is short for information value and n the total number of possible categories of X. The information value of an individual category is calculated as:

$$IV_{X_i} = (\%Bad_i - \%Good_i) * WOE_i, \qquad \textit{Equation 4}$$

with

$$WOE_i = \log\left(\frac{\%Bad_i}{\%Good_i}\right), \qquad \textit{Equation 5}$$

where WOE is short for Weight of Evidence and $\%Bad$ is calculated as:

$$\%Bad_i = \frac{nBads_i}{\sum_{i=1}^{n} nBads_i}, \qquad \textit{Equation 6}$$

which is similar to the calculation of $\%Good$. Finally, $nBads$ (the number of defaults) and $nGoods$ (the number of non-defaults) are observed, which completes the calculation. In general, the following interpretation can be attached to the outcome.

Information Value	Interpretation
< 0.02	No predictive power
0.02 to 0.1	Weak predictive power
0.1 to 0.3	Medium predictive power
0.3 to 0.5	Strong predictive power
> 0.5	Suspicious

Table 1: Interpretation of information value – from information value to predictive power.

Although the interpretations in Table 1 can help in deciding whether a variable should be included as an explanatory variable or not, these interpretations leave room for personal preference of the modeller. For example, if a variable has weak predictive power (i.e. information value between 0.02 and 0.1), but the variable is uncorrelated with any of the other explanatory variables and is a logical predictor of the dependent variable, it might still be included. Conversely, if a variable has strong predictive power, but the variable is highly correlated with another variable that has a very similar meaning (e.g. payments in arrears and months in arrears), it should be excluded.

Binning

In general, there are two ways of including categorical variables as explanatory variables. The first option is to include each distinct value as an indicator variable. The second option is to recode the variable as a discrete numeric variable, where each value is scored based on its weight of evidence or default rate. Generally, the first option is preferred, as the second option requires choosing a recoding algorithm. However, if the variables can take on many different potential values, including all these values as indicator variables significantly increases the number of explanatory variables and therefore the complexity of the model. To combat this problem, binning can be used. By binning different variable outcomes are combined into a lower number of possible outcomes, thus reducing dimensionality. The information value of a variable is often caused by only a few variables. Therefore, instead of including all outcomes as indicator variables, only those with the highest information value might be selected. It could also be that different outcomes have similar default rates, which would imply that these values could be combined into a joint indicator variable.

Trend check

Depending on the type of credit portfolio, the occurrence of defaults can be infrequent. From the perspective of financial institutions, this is fortunate; from a statistical perspective, this causes numerical uncertainty. It can therefore occur that explanatory variables are correlated with the target variable in a counterintuitive manner. To avoid modelling such effects, the following trend checks are conducted.

1. Reveal counterintuitive trends – some variables might be correlated with the dependent variable in a way that does not correspond with general business intuition. For example, it is expected that higher personal income leads to a lower

probability of default. If the observed defaults show a different trend, it is assumed that the observations are unreliable.

2. Correct outliers – some variables might display a clear trend for the majority of observations, except for a few "outliers". If there is reason to believe that these observations are not representative of the population, they could be corrected or removed from the sample.
3. Detect non-linear trends – some variables might display different trends at different intervals. For example, a trend could be increasing until a certain point and decreasing afterwards; adding a quadratic term for this variable might lead to a better fit. Other potential transformations are taking the natural logarithm, if a variable displays a non-linear but monotonically increasing or decreasing trend, or adding an interaction term, if the variable only displays a trend at specific intervals.

Correlation check

Some of the explanatory variables will be highly correlated. The correlation check is performed to select only one of the variables that are correlated, as adding both (or all, if more than two variables are correlated) will not increase the model's performance by much. Such selection is naturally performed by a selection algorithm such as the LASSO method and it could therefore be argued that there is no need for the correlation check. However, in many cases, one variable is preferred over another (for example because of data availability or common use); by manually selecting this variable, the final model is most likely to resonate with the client's model users.

Penalised regression

Once a pre-selection of potential explanatory variables has been made, further variable reduction might be accomplished by applying a statistical variable selection algorithm, such as LASSO. Such variable selection algorithms are aimed at optimising predictive power (e.g. R^2, information criteria), whilst penalising complexity (i.e. models with many explanatory variables), to prevent overfitting. While including many variables may increase in-sample model fit, the out-of-sample predictive quality of the model deteriorates. The LASSO method explicitly penalises overly complex models, by including a penalty for every explanatory variable added to the model.

Performance testing

Ultimately, the goal of the model is to correctly predict good and bad credit loans. This can be measured by the Gini coefficient, which compares the number of correctly predicted defaults against the number of incorrectly predicted ones. The Gini coefficient can range from 0 to 1, where a higher number indicates a better model. Generally speaking, a Gini over 0.50 may be considered adequate, a Gini over 0.60 good and a Gini over 0.70 excellent. However, optimal performance depends on the number of historical defaults, as a smaller number of defaults generally causes greater statistical uncertainty, which is reflected by lower model performance.