DOMINO

Sample Model Documentation

Note: The sample text in this section is taken from Logistic Regression Credit Scorecard, CC BY-NC-SA 4.0, openriskmanual.org

Definition

Credit Scorecards based on Logistic Regression are a type of credit scoring model in widespread use to support Credit Decisioning in various Consumer Finance and SME Lending businesses.

This entry serves as the Abstract Risk Model specification of a Logistic Regression Scorecard

Model Context

A population of borrowers characterised by individual features (characteristics, attributes) associated with each obligor and assumed to represent credit score factors, that is, indicators of propensity to default. The population is modelled statistically for the likelihood for defaulting (or not) over a defined period of time (the Risk Horizon).

Model Classification

The model belongs to the following categories

- discriminative (the population characteristics are not modelled)
- parametric (a set of weights represents the parameters of the model)
- exclusively observed variables
- non-linear (generalized linear)
- supervised (the historical default behaviour represents the label)
- elementary algorithm (represented by the logit model class
- frequentist approach (most commonly estimated directly on historical datasets)

Model Description

Response Variable

In binary or dichotomous logistic regression the response variable $D \in 0,1$ follows a Bernoulli distribution. The response variable D captures the realization or not of a Credit Event involving the i-th borrower D_i :

- 1, if credit event
- 0, otherwise

Explanatory Variables

The Explanatory Variables form an n dimensional vector x that comprises potentially of both continuous and discrete (categorical) variables

Model Parameters

A $n+1$ dimensional vector β (including the offset as the zero-th element) of parameters to be estimated

The Functional Form

$$p_i=\frac{\exp\beta^T x_i}{1+\exp\beta^T x_i}$$

Model Estimation

Estimation is via Maximum Likelihood. Given a vector of realized outcomes D_i the likelihood function L is:

$$L(\beta)=\prod_{i=1}^{n} p\left(x_i\right)_i^D\left(1-p\left(x_i\right)^{1-D_i}\right)$$

The log-likelihood is

$$l(\beta)=\sum_{i=1}^{n} ¿¿$$

Stylized Model Assumptions

- The observations of the response variable D are independent from each other. This assumption is manifestly not true as it is well established the defaults are correlated (Default Dependency)
- The explanatory variables combine in a linear fashion
- The specific logistic functional form is ad-hoc and does not have any economic interpretation

Some Python code and tables

Here is a demonstration of some Python code, included in the model documentation.

```
import time
# Quick, count to ten!
for i in range(10):
    # (but not *too* quick)
    time.sleep(0.5)
    print i
```

Tables can look like this:

size	material	color
9	leather	brown
10	hemp canvas	natural
11	glass	transparent