A logistic regression model for consumer default risk

Nobert Ogwel 091328, Millicent Menya 89431, Stacy Joan 145786, Kemboi Faith 204085
SIMS, Strathmore University

1 Introduction

Credit scoring is the assessment of the risk associated with lending to an organization or an individual. Credit risk modeling, namely its component Probability of Default (PD), is very helpful in the consumer credit loan grant decision. A bad customer (Defaulted) is commonly taken to be someone who has missed three consecutive months of payments. In fact, three months (or 90 days) of arrears is a standard definition of default at the international level, although it is not the only one.

Models of credit scoring are based on historical information from a dataset of existing clients, in order to assess whether the prospective client will have a greater chance of being a good or bad payer. The ability of a performance measure to capture the true skill of a model is highly dependent on the data available for assessment [4]. Beyond the social-economical characteristics of the individual, the underlying economic conditions also have a major impact on default.The existence of correlations in the data used to assess the PD invalidates using statistical tests that require an assumption of independent observations. The logistic regression model provides an appropriate statistical treatment of these correlations.

The advantages of using regression models are that it allows to perform statistical tests to identify how important are each of the application form questions to the accuracy of classification, and whether two different questions are essentially asking the same thing and getting equivalent responses. This allows to drop unimportant questions, making scorecards more robust, and helps in deciding what questions to ask in new scorecards.

In this study, a logistic regression model is applied to the UCI German credit dataset to evaluate the default risk of consumer loans.

2 Logistic Regression

When the response variable Y follows a Bernoulli distribution of parameter μ, then the generalized linear model uses the *logit function* as the canonical link function and becomes a *logistic regression model*. As Y_i $Ber(\mu_i)$ then $\mu_i = P(Y_i = 1)$.

The variable Default is a binary variable Y such that Y = 1 if defaulted, and 0 otherwise. Using the logistic regression model, the PD is a function of a set of explanatory variables X as follows:

$$P(Y = 1|X) = \frac{1}{1 + e^{-\beta X}}$$

To estimate the regression coefficients of the GLM models, the maximum likelihood method is used. The implementation provided by the command glm from R is used. The estimates for β are obtained as solution of a system of likelihood equations, that is usually solved using the Nelder and Wedderburn algorithm, which is an iterative method that uses Fisher's information matrix.

3 Data Description

The German Credit data contains financial data regarding consumer loans and a brief social characterization of the clients. It is composed of 20 variables, of which seven are quantitative and thirteen are qualitative(Metadata attached on separate file on repository)

The nonparametric Mann–Whitney–Wilcoxon test was used to compare the medians of each variable and the results show that, for a 5% significance level, there are differences between the medians in the groups of defaulters and non-defaulters, for the variables Duration in month, Credit amount, Installment-rate, and Age. Also, when testing the variables Present residence, Existing credits, and Dependents there is no statistical evidence, at a 5% significance level, to reject the null hypothesis. Hence these variables may not be relevant to explain the variable default.

Pearson Chi-squared independence test was used to check if the qualitative variables have some influence on the probability of occurring a default. The results show that the credit default risk depends on all qualitative variable except for the Job and Telephone variables.

4 Logistic regression model

For building the logistic regression model, a simple random sample of 80% of the records was considered. First, a logistic regression model was fit to the sample of 800 records,and then this model was applied to the entire original dataset, consisted of 1000 records,to predict the variable Default(Credit class).

4.1 Building a logistic regression model

Several logistic regression models for predicting the default risk were tested. For the selection of the most suitable model, the likelihood ratio test and the AIC (Akaike Information Criterion) were used.

Under the Statsmodels environment, the variables for which the null hypothesis of the Wald test is rejected, at a significance level of 5%, and therefore significant covariables in the model, are: *Duration in month, Credit amount, Installment rate, Existing checking account A12, Existing checking account A13, Existing checking account A14, Credit history A32, Credit history A33, Credit history A34, Purpose A41, Purpose A410, Purpose A42, Purpose A43, Purpose A48, Purpose A49, Savings account bonds A62, Savings account bonds A64, Savings account bonds A65, Present employment since A72, Other debtors guarantors A103, Property A122, Property A123, Property A124, Housing A152, Housing A153, Telephone A192, Foreign worker A202.*

The summary of this model is presented in Table 1 and below equation.

$$log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_n X_n$$

where μ is the mean of target variable (and represents the PD) and $X_i; (i = 1 : n)$ are the statistically significant variables Table 1.

The resulting model confirms most of the conclusions obtained in Data Description section. All the variables suggested by the exploratory analysis were found to be significant in the model.

Table 1: Summary of the logistic regression model obtained with Statsmodels environment

Coefficients	**Estimate**	**std**	**z**	**P>** $\|z\|$
const	-1.2806	1.033	-1.24	0.215
Duration_in_month	-3.4428	0.684	-5.03	0
Credit_amount	-1.9709	0.86	-2.292	0.022
Installment_rate	-1.0232	0.271	-3.782	0
Existing_checking_account_A12	0.753	0.224	3.355	0.001
Existing_checking_account_A13	1.4777	0.383	3.854	0
Existing_checking_account_A14	2.331	0.239	9.75	0
Credit_history_A32	1.4043	0.497	2.828	0.005
Credit_history_A33	2.0326	0.538	3.781	0
Credit_history_A34	2.4934	0.496	5.028	0
Purpose_A41	2.554	0.394	6.486	0
Purpose_A410	2.1529	0.817	2.636	0.008
Purpose_A42	1.1037	0.268	4.114	0
Purpose_A43	1.0635	0.255	4.178	0
Purpose_A48	2.5507	0.981	2.601	0.009
Purpose_A49	0.811	0.355	2.286	0.022
Savings_account_bonds_A62	0.6429	0.317	2.028	0.043
Savings_account_bonds_A64	1.2776	0.5	2.553	0.011
Savings_account_bonds_A65	1.6707	0.293	5.708	0
Present_employment_since_A72	-1.0231	0.479	-2.137	0.033
Other_debtors_guarantors_A103	1.6967	0.461	3.684	0
Property_A122	-0.8052	0.26	-3.102	0.002
Property_A123	-0.6197	0.24	-2.585	0.01
Property_A124	-1.6179	0.467	-3.466	0.001
Housing_A152	0.6198	0.249	2.491	0.013
Housing_A153	1.5931	0.529	3.009	0.003
Telephone_A192	0.6288	0.214	2.942	0.003
Foreign_worker_A202	2.0292	0.682	2.975	0.003

Only the variable Telephone (Telephone A192 : yes, registered under the customers name), that was not suggested by the exploratory analysis to be relevant, is now found to be relevant too.

4.2 Penalized logistic regression

In this analysis, we apply regularized logistic regression using scikit-learn's LogisticRegressionCV, which combines logistic regression with built-in cross-validation and regularization. This model is particularly useful when working with **high-dimensional data or multicollinearity**, as it applies either L1 (Lasso), L2 (Ridge), or Elastic Net penalties to shrink coefficients and reduce overfitting. The cross-validation component automatically selects the optimal regularization strength, improving model generalization and reducing the need for manual hyperparameter tuning.

This last model has a smaller AIC (309.19) than the AIC of the previous model(1043.45) and one where we Test for interactions between variables (319.83), which indicates that this model would be preferable to all our models.

The equation of this model is:

$$log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_n X_n$$

where μ is the mean of target variable (and represents the PD) and $X_i; (i = 1 : n)$ are the significant variables Table 2.

Table 2: Summary of the Penalized (lasso) logistic regression model

Significant Variable	β	$exp(\beta)$
intercept	1.192398227	3.294973835
Duration_in_month	-0.415702	0.659876884
Credit_amount	-0.287541	0.750105812
Installment_rate	-0.316537	0.72866805
Present_residence	-0.01141	0.988654847
Age	0.067633	1.069972556
Existing_credits	-0.110301	0.89556453
Dependents	-0.110121	0.895725746
Existing_checking_account_A12	0.17831	1.195195775
Existing_checking_account_A13	0.224737	1.251993399
Existing_checking_account_A14	0.785378	2.193235827
Credit_history_A31	-0.084551	0.918924789
Credit_history_A32	0.27059	1.310737558
Credit_history_A33	0.237523	1.268104163
Credit_history_A34	0.592697	1.808860339
Purpose_A41	0.529089	1.697385286
Purpose_A410	0.130796	1.139735252
Purpose_A42	0.228712	1.256979977
Purpose_A43	0.327079	1.386911039
Purpose_A46	-0.006833	0.993190292
Purpose_A48	0.153915	1.166391739
Purpose_A49	0.126434	1.134774554
Savings_account_bonds_A62	0.045242	1.046281029
Savings_account_bonds_A63	0.025924	1.02626295
Savings_account_bonds_A64	0.172113	1.187812048
Savings_account_bonds_A65	0.367468	1.444073587
Present_employment_since_A72	-0.211555	0.809324767
Present_employment_since_A73	-0.123283	0.884013451
Present_employment_since_A74	0.115544	1.122483903
Marital_and_sex_A93	0.157343	1.170396991
Marital_and_sex_A94	0.011934	1.012005494
Other_debtors_guarantors_A102	-0.122678	0.884548441
Other_debtors_guarantors_A103	0.232502	1.26175297
Property_A122	-0.151771	0.859185012
Property_A123	-0.156367	0.855245258
Property_A124	-0.294455	0.744937468
Other_installment_plans_A142	-0.000153	0.999847012
Other_installment_plans_A143	0.150575	1.16250249
Housing_A152	0.201594	1.223351227
Housing_A153	0.241501	1.273158728
Job_A172	-0.002928	0.997076282
Job_A174	-0.093108	0.911095097
Telephone_A192	0.153621	1.16604887
Foreign_worker_A202	0.21547	1.240444769

4.3 Model Estimates

In Table 2, the estimates for the model parameters are shown.

In logistic regression models, rather than looking at the coefficients β_i, it is more important to focus on the values of $\exp(\beta_i)$, because they represent the influence that the increase in an independent variable X_i has in the probability of the dependent variable Y becoming 1.

It follows that:

$$log\left(\frac{P(Y=1|X_i)}{P(Y=0|X_i)}\right) = \beta_0 + \beta_1 X_1 + ...\beta_i X_i + ...\beta_p X_p$$

$$\frac{P(Y=1|X_i)}{P(Y=0|X_i)} = e^{\beta_0+\beta_1 X_1+...\beta_i X_i+...\beta_p X_p}$$

The term on the left side of above equation is called the *odds* of the variable Y. In our model, it represents the ratio between the probability of a client committing default and the probability of not committing default.

$exp(\beta_i) = (Odds(Y|X_i + 1))/Odds(Y|X_i)$ represents the OR. The estimates of the coefficient β_i of our optimal logistic regression model are presented in Table 2

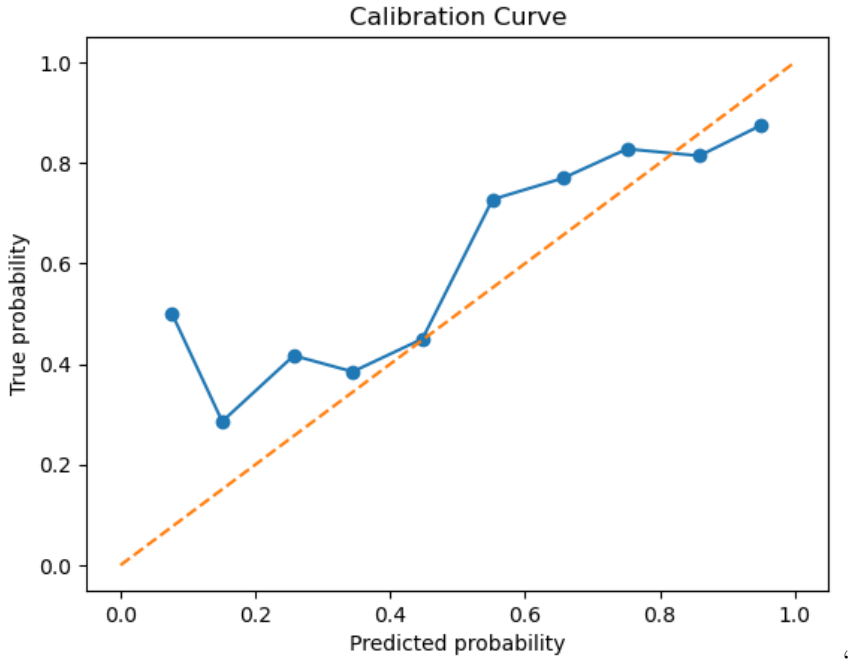
For instance, the variable Age has an odds ratio of approximately 1.07, meaning that for each additional year in a borrower's age, the odds of default increase by about 7%. In contrast, Duration in month has an odds ratio of 0.66, implying that longer loan durations are associated with a 34% decrease in the odds of default. Similarly, higher Credit amount and Installment rate are associated with lower chances of default, with odds decreasing by roughly 25% and 27% respectively for each unit increase. Present residence shows a nearly neutral effect, with its odds ratio close to 1, indicating it has little impact on default probability. These interpretations provide valuable insights into how different borrower and loan characteristics influence credit risk.

5 Model validation

Before using our optimal model to estimate the probability of a client of the bank defaulting, the model has to be validated through a series of statistical tests, and the assumptions of the model have to be verified

5.1 Goodness-of-fit tests

An important topic in modeling exercise is the goodness-of-fit test: testing the null hypothesis that the model fits the data well versus the opposite.

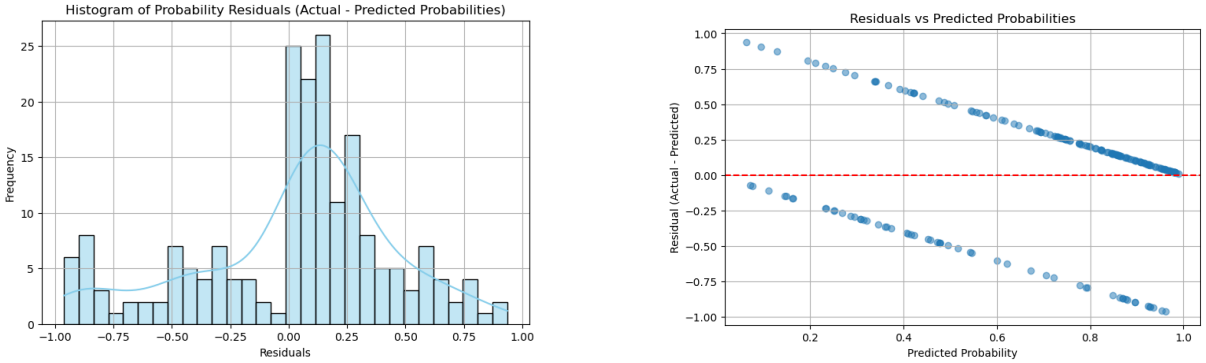


The logistic regression model demonstrates a reasonable but imperfect goodness-of-fit. The calibration curve indicates that the model overestimates the probability of the positive class at lower predicted probability ranges (0.0–0.4), where the observed true outcomes are significantly lower than the predicted

values. In contrast, the model performs better in the mid to higher probability ranges (above 0.5), although it tends to slightly underestimate the likelihood of positive outcomes in these regions. Despite these calibration issues, the model shows a generally increasing relationship between predicted probabilities and observed outcomes, suggesting it maintains adequate discriminatory power.

5.2 Residuals analysis

The model may also be validated by studying the residuals and performing regression diagnostics.



(a) Histogram of Probability Residuals

(b) Residuals vs Predicted Probabilities

Figure 1: Residuals analysis Plots

The diagnostic plots of the logistic regression model reveal notable patterns that suggest potential issues related to data size and calibration. The histogram of residuals (Actual - Predicted Probabilities) shows a distribution centered near zero, indicating a generally reasonable model fit. However, the presence of asymmetry and multiple peaks suggests potential calibration issues, likely intensified by a small dataset. The residuals versus predicted probabilities plot exhibits the expected linear bands for binary classification, but the concentration of predictions near the extremes and the structured spread of residuals further support the possibility of overfitting or limited generalization. These observations imply that the model's predicted probabilities may not be well-calibrated, and that a larger or more representative dataset could improve stability and performance.

6 Conclusions