**Report**

Context

Banks play a crucial role in market economies. They decide who can get finance and on what terms and can make or break investment decisions. For markets and society to function, individuals and companies need access to credit.

Credit scoring algorithms, which make a guess at the probability of default, are the method banks use to determine whether or not a loan should be granted. This project aim to improve on the state of the art in credit scoring, by *predicting the probability that somebody will experience financial distress in the next two years*.

Objectives

General

* Improve credit scoring
* Build a model that borrowers and lenders can use to make better financial decisions

Concrete

* Generate delinquency scores for the data provided

Methodology

Given the data provided, the following approach was implemented to get our results.

1. Exploration of data: visual identification of patterns, trends, and extreme values in data
2. Modelling proposal: identification of logistic regression as the model that would potentially yield the best results, given objectives and current data
3. Processing data: imputation of data on empty datapoints with median values
4. Estimating the model: we propose and build three models, then compare the best fit.
5. Evaluating the model: we use accuracy as we believe it is the best measure for evaluating binary responses

Results

The three tested models yielded similar results. The best trained model has an accuracy of 0.9338. Nevertheless, the other two models performed almost as well. This model was used with the data on the file ‘cs-test.csv’. Results are stored on “Fitted.txt”.

Next steps

* Further iterations of this exercise could include a Linear Probability Model and a Probit regression as different classifiers
* Crosstabs suggest the need to bin variables into groups, treating them as dummy variables. These could yield better predictions of our data.
* Evaluation should include other measures: mean square error, mean absolute deviation, and entropy among others

**Technical appendix**

|  |  |  |
| --- | --- | --- |
| **Description** | **Original header** | **Relabeling** |
| (None) | (No header) | ‘ID’ |
| Has the person experienced past 90 days due delinquency or worse? | 'SeriousDlqin2yrs' | ‘SD2Y’ |
| Total balance on credit cards and personal lines of credit[[1]](#footnote-1) divided by the sum of credit limits | 'RevolvingUtilization  OfUnsecuredLines' | ‘RUUL’ |
| Age of borrower in years | 'age' | ‘Age’ |
| Times borrower has been 30-59 days past due but no worse in the last 2 years | 'NumberOfTime30-59  DaysPastDueNotWorse' | ‘LP30\_59’ |
| Monthly debt payments, alimony, and living costs, divided by monthy gross income | 'DebtRatio' | ‘DR’ |
| Monthly income | 'MonthlyIncome' | ‘MI’ |
| Number of open loans[[2]](#footnote-2) and lines of credit[[3]](#footnote-3) | 'NumberOfOpen  CreditLinesAndLoans' | ‘OCLL’ |
| Number of times borrower has been 90 days or more past due in the last 2 years | 'NumberOfTimes  90DaysLate' | ‘LP90\_’ |
| Times borrower has been 60-89 days past due but no worse in the last 2 years | 'NumberRealEstate  LoansOrLines' | ‘LP60\_90’ |
| Number of mortgage and real estate loans[[4]](#footnote-4) | 'NumberOfTime60-89  DaysPastDueNotWorse' | ‘MREL’ |
| Number of dependents, excluding themselves | 'NumberOfDependents' | ‘Deps’ |

**Table 1. Data provided**

Notes on methodology

1. Exploration of data

Running the function go.py will yield several information on data:

* Descriptive statistics, to get an idea of the distributions of data for each variable
* Histograms, which resulted difficult to read given the numerous ‘extreme’ observations on variables (e.g. MI > 3,000,000, DR > 5,000)
* Scatter plots, of variables vs. observation number, to identify potential bins to categorize or group data [[5]](#footnote-5)

1. Modelling proposal

We use logistic regression, as we consider that it is a natural model to fit probabilities.

ID SD2Y RUUL Age LP30\_59 DR

count 150000.000 150000.000 150000.000 150000.000 150000.000 150000.000

mean 75000.500 0.067 6.048 52.295 0.421 353.005

std 43301.415 0.250 249.755 14.772 4.193 2037.819

min 1.000 0.000 0.000 0.000 0.000 0.000

25% 37500.750 0.000 0.030 41.000 0.000 0.175

50% 75000.500 0.000 0.154 52.000 0.000 0.367

75% 112500.250 0.000 0.559 63.000 0.000 0.868

max 150000.000 1.000 50708.000 109.000 98.000 329664.000

MI OCLL LP90\_ LP60\_90 MREL Deps

count 1.203e+05 150000.000 150000.000 150000.000 150000.000 146076.000

mean 6.670e+03 8.453 0.266 1.018 0.240 0.757

std 1.438e+04 5.146 4.169 1.130 4.155 1.115

min 0.000e+00 0.000 0.000 0.000 0.000 0.000

25% 3.400e+03 5.000 0.000 0.000 0.000 0.000

50% 5.400e+03 8.000 0.000 1.000 0.000 0.000

75% 8.249e+03 11.000 0.000 2.000 0.000 1.000

max 3.009e+06 58.000 98.000 54.000 98.000 20.000

**Table 2. Descriptive statistics**

1. Processing data

We found missing information in observations of two variables, which were filled with their corresponding median values:

* Monthly income
* Dependents in family, excluding oneself

1. Estimating the model and classifying

The three models are estimated without binning or discretizing any variables.

* The first model regresses SD2Y with the rest of the covariates (except ID). The classification rule of the logistic regression is as follows:

Results: Logit

=================================================================

Model: Logit Pseudo R-squared: 0.069

Dependent Variable: SD2Y AIC: 68577.0382

Date: 2016-04-12 23:07 BIC: 68676.2221

No. Observations: 150000 Log-Likelihood: -34279.

Df Model: 9 LL-Null: -36808.

Df Residuals: 149990 LLR p-value: 0.0000

Converged: 1.0000 Scale: 1.0000

No. Iterations: 7.0000

------------------------------------------------------------------

Coef. Std.Err. z P>|z| [0.025 0.975]

------------------------------------------------------------------

RUUL -0.0001 0.0001 -0.8887 0.3742 -0.0002 0.0001

Age -0.0503 0.0005 -94.5485 0.0000 -0.0514 -0.0493

LP30\_59 0.4913 0.0112 43.8688 0.0000 0.4693 0.5132

DR -0.0000 0.0000 -3.5100 0.0004 -0.0001 -0.0000

MI -0.0001 0.0000 -16.3768 0.0000 -0.0001 -0.0000

OCLL -0.0207 0.0026 -8.0156 0.0000 -0.0258 -0.0157

LP90\_ 0.4196 0.0149 28.0714 0.0000 0.3903 0.4489

LP60\_90 0.1034 0.0107 9.7043 0.0000 0.0825 0.1243

MREL -0.8834 0.0175 -50.4487 0.0000 -0.9178 -0.8491

Deps 0.0383 0.0090 4.2792 0.0000 0.0208 0.0559

=================================================================

* The second model takes out the RUUL, as it results statistically insignificant. The classification rule of the logistic regression remains unchanged.

Results: Logit

=================================================================

Model: Logit Pseudo R-squared: 0.069

Dependent Variable: SD2Y AIC: 68576.0321

Date: 2016-04-12 23:07 BIC: 68665.2976

No. Observations: 150000 Log-Likelihood: -34279.

Df Model: 8 LL-Null: -36808.

Df Residuals: 149991 LLR p-value: 0.0000

Converged: 1.0000 Scale: 1.0000

No. Iterations: 7.0000

------------------------------------------------------------------

Coef. Std.Err. z P>|z| [0.025 0.975]

------------------------------------------------------------------

Age -0.0503 0.0005 -94.5529 0.0000 -0.0514 -0.0493

LP30\_59 0.4913 0.0112 43.8714 0.0000 0.4693 0.5132

DR -0.0000 0.0000 -3.5176 0.0004 -0.0001 -0.0000

MI -0.0001 0.0000 -16.3962 0.0000 -0.0001 -0.0000

OCLL -0.0207 0.0026 -8.0025 0.0000 -0.0258 -0.0156

LP90\_ 0.4196 0.0149 28.0726 0.0000 0.3903 0.4489

LP60\_90 0.1033 0.0107 9.6965 0.0000 0.0825 0.1242

MREL -0.8835 0.0175 -50.4516 0.0000 -0.9178 -0.8492

Deps 0.0383 0.0090 4.2779 0.0000 0.0208 0.0559

=================================================================

* The third model keeps the covariates of the second model, but elevates the threshold of the classification rule to *u = 0.9.*

Results: Logit

=================================================================

Model: Logit Pseudo R-squared: 0.069

Dependent Variable: SD2Y AIC: 68576.0321

Date: 2016-04-12 23:07 BIC: 68665.2976

No. Observations: 150000 Log-Likelihood: -34279.

Df Model: 8 LL-Null: -36808.

Df Residuals: 149991 LLR p-value: 0.0000

Converged: 1.0000 Scale: 1.0000

No. Iterations: 7.0000

------------------------------------------------------------------

Coef. Std.Err. z P>|z| [0.025 0.975]

------------------------------------------------------------------

Age -0.0503 0.0005 -94.5529 0.0000 -0.0514 -0.0493

LP30\_59 0.4913 0.0112 43.8714 0.0000 0.4693 0.5132

DR -0.0000 0.0000 -3.5176 0.0004 -0.0001 -0.0000

MI -0.0001 0.0000 -16.3962 0.0000 -0.0001 -0.0000

OCLL -0.0207 0.0026 -8.0025 0.0000 -0.0258 -0.0156

LP90\_ 0.4196 0.0149 28.0726 0.0000 0.3903 0.4489

LP60\_90 0.1033 0.0107 9.6965 0.0000 0.0825 0.1242

MREL -0.8835 0.0175 -50.4516 0.0000 -0.9178 -0.8492

Deps 0.0383 0.0090 4.2779 0.0000 0.0208 0.0559

=================================================================

1. Evaluating the model

* Accuracy model 1: 0.9337533333333333
* Accuracy model 2: 0.9337533333333333
* Accuracy model 3: 0.93318

We selected the second model as it has the least covariates and best accuracy of the tested cases.

1. Except real estate and no installment debt [↑](#footnote-ref-1)
2. e.g. installment such as car loan or mortgage [↑](#footnote-ref-2)
3. e.g. credit cards [↑](#footnote-ref-3)
4. Including home equity lines of credit [↑](#footnote-ref-4)
5. Both histograms and scatter plots are avaliable in their corresponding folder for consultation [↑](#footnote-ref-5)