Test Case Solution

by Andrii Osipa for Econometrix Contest 2015

Data

Lets have a look at given dataset Training Sample.

First remark: data must describe real situation, and according to the nature of variables they must conform with some natural rules.

* income : ≥0;

* expenditures : ≥0;

* loan amount : ≥0;

* collateral amount : ≥0;

* age: ≥0 and have some natural upper bound, for example, 130;

* working experience: ≥0 and ≤(age-18).

Example of incorrect data: Client with ID t4 is 214 years old and has -3 years of working experience. Such data must be excluded from the datasets.

Correlation Table(Pearson)							
	Customer's monthly income	Customer's monthly expenditures	Requested loan amount	Monthly loan installment	Collateral value	Customer's age	Customer's work experience
Customer's monthly income	1	**0,67756843**	0,209538873	**0,589823903**	0,135873391	0,05698615	0,269644953
Customer's monthly expenditures	0,67756843	1	0,250219554	0,336353868	0,107558194	-0,003059068	0,162138276
Requested loan amount	0,209538873	0,250219554	1	0,024975861	0,298696804	0,106615019	0,104410882
Monthly loan installment	0,589823903	0,336353868	0,024975861	1	-0,053738216	0,038505895	0,065997623
Collateral value	0,135873391	0,107558194	0,298696804	-0,053738216	1	-0,038979682	-0,043103655
Customer's age	0,05698615	-0,003059068	0,106615019	0,038505895	-0,038979682	1	0,20047211
Customer's work experience	0,269644953	0,162138276	0,104410882	0,065997623	-0,043103655	0,20047211	1

Conclusions:

- Income and expenditures are, probably, not independent.
- Income and monthly loan installment are, probably, not independent.

Problem with highly correlated variables: it makes estimation of regression parameters less accurate and the point in this case is finding the most exact values of the parameters.

New Variables that are used to predict default

OverExpenditures parameter.

It is known that expenditures(Y) depend on income(X) in such way Y = a + b*X.

For a loaner money paid as a loan installment(Z) also become an every month expenditures, so it is logical to say that total expenditures(Y2) is described with the formula Y2 = Y + Z. Then overExpenditures(S) is defined by S = Y2 – a - b*X, where we have Y2 – real expenditures and (a + b*X) is "standard" expenditures for someone with income X. Moreover, variables S and X are independent now.

Correlation coefficient of two variables (expenditures + loan installment) and (income) is 0,776, so there is probably a linear dependence between them.

New Variables that are used to predict default

Collateral amount and loan amount.

Generally it is not important how expensive property that is collateral(C) for a loan while it is bigger then the amount of loan(A). So, value C/A is more important than real amounts of money. Further two basic variables **collateral amount** and **loan amount** are replaced by new variable C/A.

If $C \geq A$ then in case of default bank will not have any losses in case of default. We can also interpret this situation in such way $(C/A) \geq 1$.

In other case bank will have a loss and it is C-A. This case is defined by that fact that A/C is less than 1, close to 0.

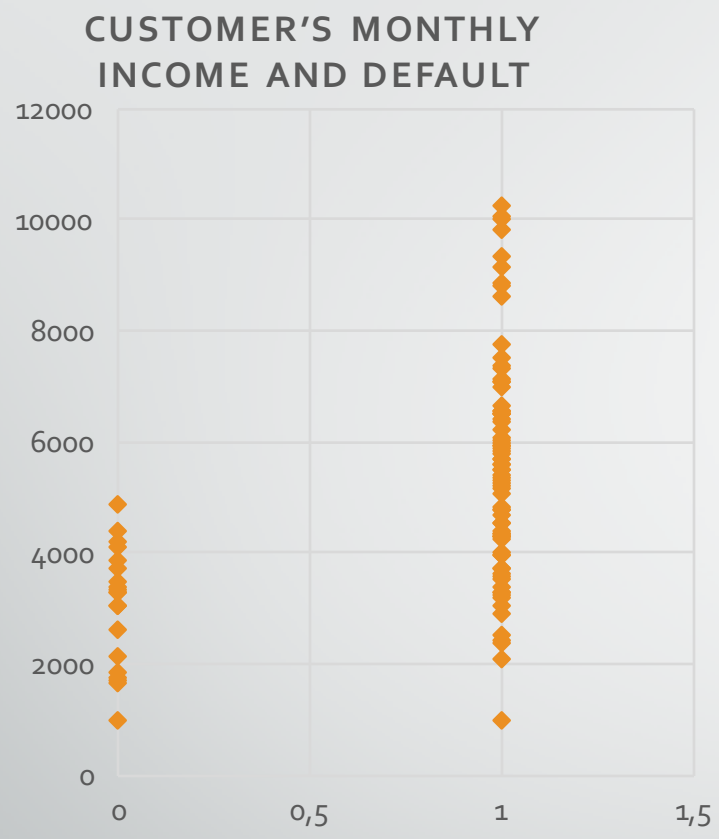
Correlation table for new variables

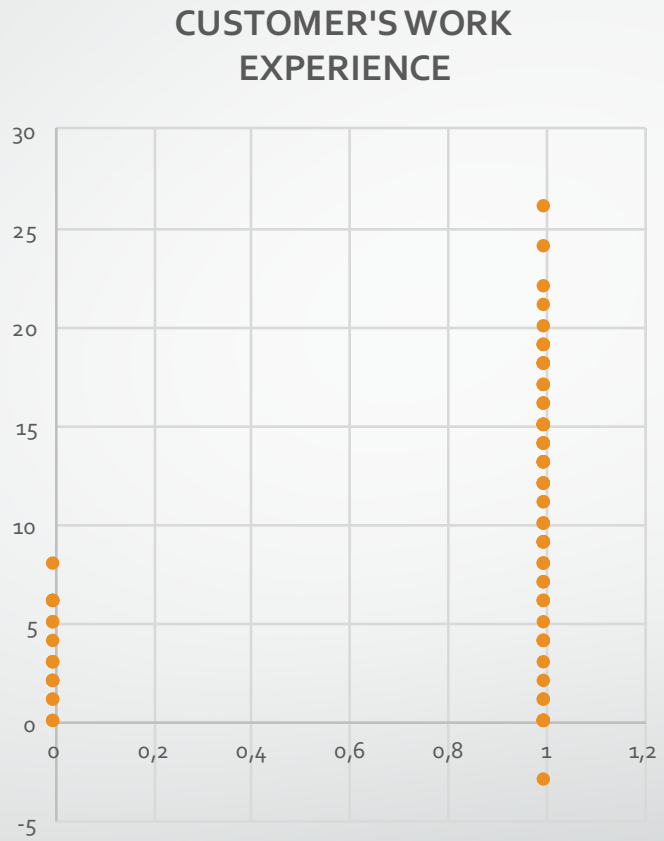
	Customer's monthly income	Customer's age	Customer's work experience	Collateral / loan amount	overExpences
Customer's monthly income	1,0000000	0,0569861	**0,2696450**	0,0730756	0,0000000
Customer's age	0,0569861	1,0000000	0,2004721	-0,0421510	-0,0369725
Customer's work experience	0,2696450	0,2004721	1,0000000	-0,0157084	-0,1083051
Collateral / loan amount	0,0730756	-0,0421510	-0,0157084	1,0000000	-0,1250254
overExpences	0,0000000	-0,0369725	-0,1083051	-0,1250254	1,0000000

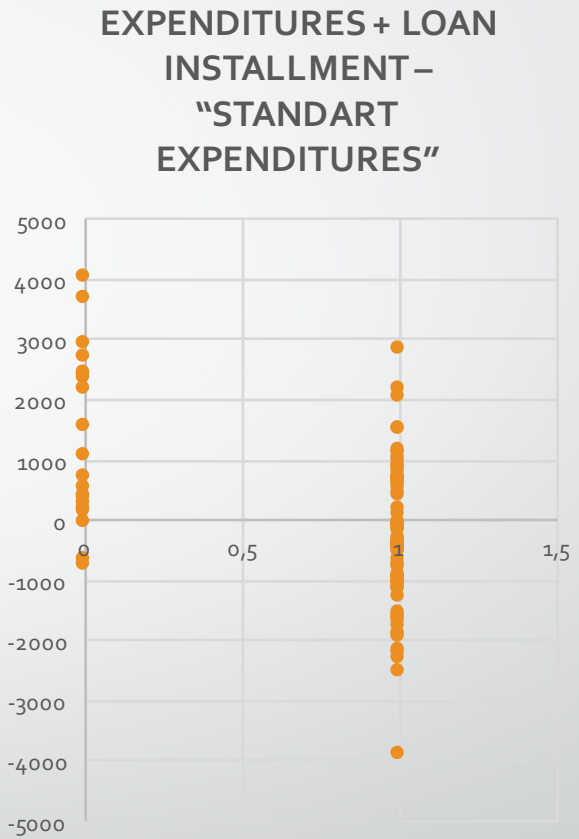
All coefficients here are quite close to zero, so we can assume that all variables in this set are independent.

Using highly correlated variables is not effective in this case because it will cause that the estimated values of parameters will be very different from real values of these parameters. Therefore the prediction function will be very different from real one and will not give accurate result.

Univariate Analysis







In this case univariate analysis is provided visually. On the scatter plots it is easy to see if the parameter is separating the data effectively into two categories : Default(0) and No-Default(1). If it is a good separator it means that it is quite important factor in general model. The most important factors, according to this type of analysis, are provided here.

Regression model & other models

Variables	Customer's monthly income	Customer's age	Customer's work experience	Collateral / loan amount	overExpences
	x1	x2	x3	x4	x5

Coefficients	
Intercept	0,2045066
X Variable 1	0,0000710
X Variable 2	-0,0001282
X Variable 3	0,0199247
X Variable 4	0,0280954
X Variable 5	-0,0001221

$$\boldsymbol{Default = 0,2045 + 0,0000710 * x1 - 0,00012 * x2 + 0,0199 * x3 + 0,0280 * x4 - 0,0001221 * x5}$$

Result interpretation: if Default value is greater than 0.5 then we assume that it is 1, otherwise it is 0.

Best model is selected by defining accurately confidence rate and training & testing model which depends on different sets of variables. If confidence rate is too close to 100% than model may be overfitted and will describe errors but not the nature of the process.

Other model, that can be accurate enough in this case, is decision tree. It works on the set good enough because there are not many parameters and these parameters are separating space of data in two parts with enough accuracy.

Results

Regression, which depends on variables **income**, **working experience**, **age**, **overExpences** and **ratio of collateral amount to loan amount**, gives accuracy of **97%** on Validation Sample. And all errors are False Positive errors: model gives answer 1 when real is 0.

Testing the model on different samples is important because it was trained on a small set of data and in real life data may be very different because basic variables are positive integer numbers and some of them have no upper bound, for example. Therefore we can invent artificially many very different cases and it is important how model will be behave on them.