**Lessons Learned Report**

**FOR**

**Guido Rossum**

**BY**

**Clark Humphrey**

**Introduction:**

This report describes my key take-aways and lessons learned from preparing and exploring the data for determining if we can ensure that customers can/will pay their loans. I’m also utilizing Python and Jupyter notebooks for this project, so their use is also part of the scope of this report.

**Feature Selection and Preparation:**

The data set explored is described in documentation provided as:

* “This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:
* X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
* X2: Gender (1 = male; 2 = female).
* X3: Education (1 = graduate school; 2 = university; 3 = high school; 0, 4, 5, 6 = others).
* X4: Marital status (1 = married; 2 = single; 3 = divorce; 0=others).
* X5: Age (year).
* X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -2: No consumption; -1: Paid in full; 0: The use of revolving credit; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
* X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
* X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .; X23 = amount paid in April, 2005. Y: client's behavior; Y=0 then not default, Y=1 then default"

In addition, the data file has an “ID” variable that numbers each record consecutively from 1 on. Exploration of the data revealed it has 30,000 records and no fields were “null” or “NA”.

The following was done to select the features and prepare the data for analysis:

1. X3: Since this field has multiple values for “others”, all values of “0”, “5,” and “6” were changed to “4” so only a single value (“4”) represents “others”.
2. The response variable field was originally named “default payment next month” which includes spaces. In order to not potentially cause any interference with processing models, the name was changed to “default\_next\_month” so no spaces are present.
3. Many of the features are of nominal data. Research revealed that pandas has a “category” type for nominal data. The following fields were changed to “category” type: “default\_next\_month”, X2 (Gender), X3 (Education), X4 (Marital status), and X6-X11 (History of payment).
4. The feature “ID” was removed since it’s not needed for analysis.

**What worked well…**

* Working with Python and its extensions (pandas, matplotlib, etc.) went well. The high amount of documentation accessible on the internet has proven so far to be helpful in finding Python coding solutions for any situations I’ve encountered, such as how to change an integer data type to an ordinal data type.
* Working with Jupyter notebooks is starting to click. There are parallels to how R Studio works, but the way cells are used and that the output generated by the cell’s code is listed directly below the cell is very nice.

**What didn’t work well…**

* In comparing Jupyter notebooks to R Studio, I miss some of the information contained in some of the windows in R Studio, such as information about the data frames that have been imported and assigned.

**What I learned from this analysis that has potential business value…**

* In comparing the histograms of the “Pay\_6” and the “Pay\_0” variables, it shows a general shift over that 6-month period from practically no loans with a 1-month payment delay, to about 4,000 with such a delay. Also, the number of loans “paid in full” or with greater than 1-month payment delay remained about the same, while the number of loans with “no consumption” dropped by almost 50% (from ~5,000 to ~2,800). Using data analysis algorithms to find patterns that can predict this type of shift when it leads to a default should be very helpful to the business.
* From the box plots, most of the loans “limit balance” are less than $300,000 with “Bill\_Amt1” less than $100,000. However, the “max” value of both of these are ~$1,000,000.

**Recommendations based on these findings…**

* Proceed to the next step in the data analysis process and use algorithms to analyze the data so predictions can be made when defaults are likely in order to help the business.