Data Analytics Project 1
Credit Default Prediction Model

Credit scoring is the practice of analyzing a person's background and credit application to assess the creditworthiness of the person. One can take numerous approaches on analyzing this creditworthiness. In the end it basically comes down to first selecting the correct independent variables (e.g. income, age, gender) that lead to a given level of creditworthiness. In other words: $creditworthiness = f(income, age, gender, ...)$. A credit scoring system can be represented by linear regression, logistic regression, machine learning or a combination of these. The dataset, $creditset$, contains information on different clients who received a loan at least 10 years ago. The variables $income$ (yearly), age, $loan$ (amount in euros) and LTI (the loan to yearly income ratio) are available. I have simulated and added nine additional variables in $creditset2$:

cred_rate	Credit rating assigned as an average of ratings from three different credit bureaus.
rev_cred	Revolving credit assessment: Excessive, Moderate, Average, Minimal
health	Health score: 6 = excellent; 5 = good; 4 = moderate; 3 = average; 2 = poor; 1 = terminally ill
home_val	Home value (current market) in Euros
risk	Risk factor qualitative assessment: high, average, or low
ment_scr	Mental health score: 6 = excellent; 5 = good; 4 = average; 3 = high functioning autism; 2 = assistance required; 1 = dysfunctional
married	Current marital status: Married, Single (never been married), Divorced, or Widowed
nbr_child	Number of children (integer): 1, 2, …, n
soc-media	Social media account (active): yes or no

Your goal is to devise a model which predicts, based on the input variables, whether a default will occur within 10 years. Traditionally, age and LTI are considered the best indicators, but we want a model that will "learn" from the additional data in $creditset2$.

Requirements:

A **white-paper** report (Google what this consist of) with a maximum of five pages. You may add an appendix containing your detailed analysis results (no more than 10 additional pages). An appendix with your model code must be attached. This will give you the foundation of a conference paper.

A **PowerPoint slide deck** consisting of a maximum of about 10 slides :
1. Title slide
2. Business problem statement

3. Modeling objective
4. Model functional form and reasoning for its selection
5. Final model (variables with coefficients and explanation) summary
6. 1 deep-dive slide per model variable (a good model may only have three or four factors, even if all variables are statistically significant—modeling is an "art").

Each slide in a PowerPoint slide deck needs to stand alone yet have few words and contain one graphic (chart, graph, table, picture, etc.) with words, but not verbose. Each slide should tell a story that an executive will understand. If you cannot effectively communicate your analysis, it is worthless—a model is only good if it is used.

Stages of Model Development
1. Business problem development and model requirement definition
2. Model design plan, including modeling objective and model functional from, plus potential data sources
3. Model development
   a. Data preprocessing
   b. Exploratory analysis
   c. Model training
   d. Model testing
4. Model post processing analysis
5. Model validation
6. Model Implementation

Things to consider while performing exploratory analysis:
1. Is a high LTI score better or worse than a low LTI score?
2. Is there a direct relationship between current home market value and income? If so, what is it? If not, why not?
3. Does the qualitative risk factor generally agree with the credit scores?
4. What is the relationship between marital status and the other variables if any?
5. What is the relationship between number of children and the other variables if any?
6. Continue to ask the data these kinds of questions.

Citations: Use ISO 690 with Numerical Reference.