**Analytics Accelerator Program Capstone Project**

Analysis of German Credit Data



Data mining is a critical step in knowledge discovery involving theories, methodologies and tools for revealing patterns in data. It is important to understand the rationale behind the methods so that tools and methods have appropriate fit with the data and the objective of pattern recognition. There may be several options for tools available for a data set.

When a bank receives a loan application, based on the applicant’s profile the bank has to make a decision regarding whether to go ahead with the loan approval or not. Two types of risks are associated with the bank’s decision –

* If the applicant is a good credit risk, i.e. is likely to repay the loan, then not approving the loan to the person results in a loss of business to the bank
* If the applicant is a bad credit risk, i.e. is not likely to repay the loan, then approving the loan to the person results in a financial loss to the bank

Objective of Analysis:

**Minimization of risk and maximization of profit on behalf of the bank.**

To minimize loss from the bank’s perspective, the bank needs a decision rule regarding who to give approval of the loan and who not to. An applicant’s demographic and socio-economic profiles are considered by loan managers before a decision is taken regarding his/her loan application.

The German Credit Data contains data on 20 variables and the classification whether an applicant is considered a Good or a Bad credit risk for the loan applicants. Here is a link to the German Credit data (*right-click and "save as"* ).  A predictive model developed on this data is expected to provide a bank manager guidance for making a decision whether to approve a loan to a prospective applicant based on his/her profiles. ( Source: <https://onlinecourses.science.psu.edu>)

Required Activities

**Step # 1-** Load this data in the R environment. Data link and Dictionary can be found below-



**Step # 2-** Explore the data using different R functions such as dim (), head (), tail (), summary () etc.

**Step # 3**- Create a dependent variable using the “Default\_On\_Payment” variable of the dataset. A sample R code for which is given below-



**Step # 4-** Do data cleaning, in terms of missing values etc.

**Step # 5-** Subset the data to create 60% & 40% split in the data. Use 60% sample for developing models. Feel free to use the below code ( here the name of the original dataset is Default\_On\_Payment)-

d60percent <- Default\_On\_Payment[sample(nrow(Default\_On\_Payment), replace= F, size=0.60\*nrow(Default\_On\_Payment)),]

**Step # 6-** Do bivariate plots, WOE, IV etc. to identify key independent factors. Feel free to change variable types to numeric, categorical etc. as you see fit.

**Step # 7-** Create the right levels of bins for each significant variable (as required) and build a logistic regression model

**Step # 8-** Summarize your model performance in the following statistics-

1. Top 10 variables which explain the default behaviour
2. ROCR chart
3. Area under the curve (AUC)
4. Deciles by predicted probability of default
5. KS
6. Lift
7. Lorenz and Gini Index
8. Any other metrics that you think will be useful to explain your model’s performance

Other Helpful Documents for this Activity-

 

GOOD LUCK!