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

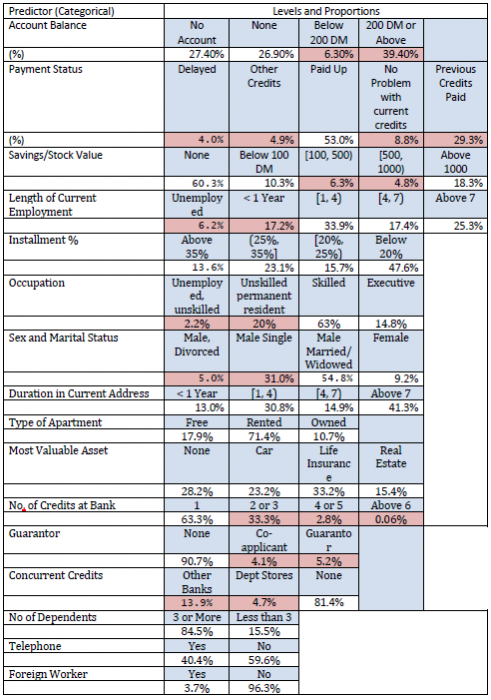
Objective of Analysis:

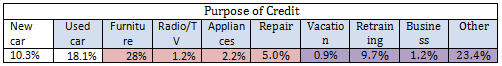
**Minimization of risk and maximization of profit on behalf of the bank.**

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 1000 loan applicants.

Proportions of applicants belonging to each classification of a categorical variable are shown in the following table (below). The pink shadings indicate that these levels have too few observations and the levels are merged for final analysis.





The following analytical approaches are taken:

* Logistic regression: The response is binary (Good credit risk or Bad) and several predictors are available.
* Discriminant Analysis:
* Tree-based method and Random Forest

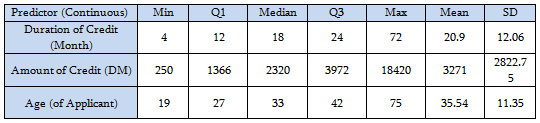
Depending on the cell proportions given in the one-way table above two or more cells are merged for several categorical predictors. We present below the final classification for the predictors that may potentially have any influence on Creditability

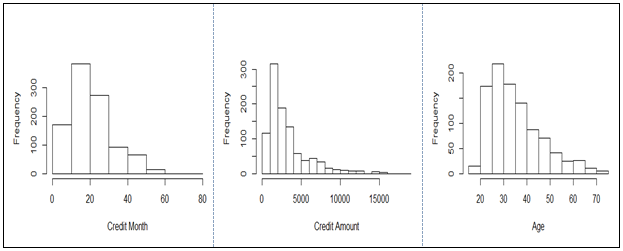
* Account Balance: No account (1), None (No balance) (2), Some Balance (3)
* Payment Status: Some Problems (1), Paid Up (2), No Problems (in this bank) (3)
* Savings/Stock Value: None, Below 100 DM, [100, 1000] DM, Above 1000 DM
* Employment Length: Below 1 year (including unemployed), [1, 4), [4, 7), Above 7
* Sex/Marital Status: Male Divorced/Single, Male Married/Widowed, Female
* No of Credits at this bank: 1, More than 1
* Guarantor: None, Yes
* Concurrent Credits: Other Banks or Dept Stores, None
* ForeignWorker variable may be dropped from the study
* Purpose of Credit: New car, Used car, Home Related, Other

Observations/Explorations:

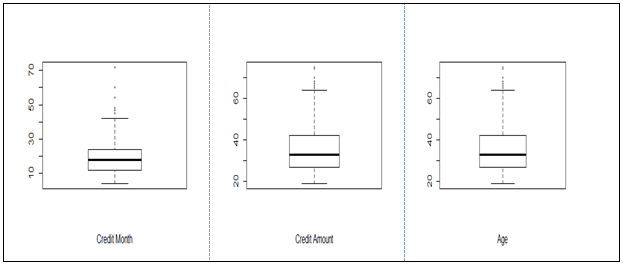
30% of 1000 applicants have no account and another 30% have no balance while 40% have some balance in their account. Among those who have no account 135 are found to be Creditable and 139 are found to be Non-Creditable. In the group with no balance in their account, 40% were found to be on-Creditable whereas in the group having some balance only 1% are found to be Non-Creditable.

Summary for the continuous variables:





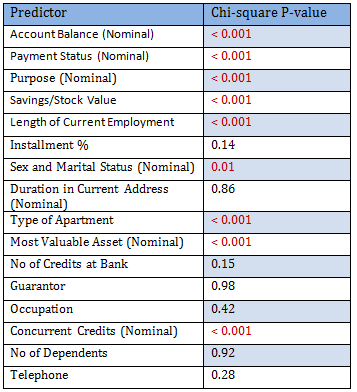
All the three variables show marked positive skewness.

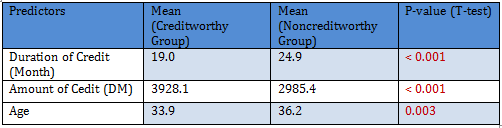


In preparation of predictors to use in building a logistic regression model, we consider bivariate association of the response (Creditability) with the categorical predictors.

**Build Logistic Regression Model**

Since the number of predictors in this problem is not very high, it is possible to look into the dependency of the response (Creditability) on each of them individually. The following table summarizes the chi-square p-values for each contingency table.

Among the sample of size 1000, 700 were Creditable and 300 Non-Creditable. This classification is based on the Bank’s opinion on the actual applicants. 



Only significant predictors are to be included in the logistic regression model. Since there are 1000 observations 50:50 cross-validation scheme is tried

1000 observations are randomly partitioned into two equal sized subsets – Training and Test data. A logistic model is fit to the Training set.

Build Three models with various parameters and check Residual deviance and AIC

The Akaike Information Criterion (**AIC**) provides a method for assessing the quality of your **model** through comparison of related models. It's based on the Deviance, but penalizes you for making the **model** more complicated. Much like adjusted **R**-squared, its intent is to prevent you from including irrelevant predictors.