A short primer of underwriting in the credit industry -

In general, whenever an individual/corporation applies for a loan from a bank (or any loan issuer), their credit history undergoes a rigorous check to ensure that whether they are capable enough to pay off the loan (in this industry it is referred to as credit-worthiness).

The issuers have a set of model/s and rule/s in place which take information regarding their current financial standing, previous credit history and some other variables as input and output a metric which gives a measure of the risk that the issuer will potentially take on issuing the loan. The measure is generally in the form of a probability and is the risk that the person will default on their loan (called the probability of default) in the future.

Based on the amount of risk that the issuer is willing to take (plus some other factors) they decide on a cutoff of that score and use it to take a decision regarding whether to pass the loan or not. This is a way of managing credit risk. The whole process collectively is referred to as underwriting.

NOTE: What I have described above is a very simplified version of the process to give you an introductory overview. You can refer to the following resources for a better understanding:

http://budgeting.thenest.com/mean-loan-goes-underwriting-23201.html

http://www.investopedia.com/ (a great source to find meanings of BFSI terminology and jargon)

Overview of the problem

In this project you will have to put yourself in the shoes of a loan issuer and manage credit risk by using the past data and deciding whom to give the loan to in the future. The text files contain complete loan data for all loans issued by XYZ Corp. through 2007-2015. The data contains the indicator of default, payment information, credit history, etc.

The data should be divided into train (June 2007 - May 2015) and out-of-time test (June 2015 - Dec 2015) data. You will have use the training data to build models/analytical solution and finally apply it to test data to measure the performance and robustness of the models.

You should use the variable : 'issue_d' to divide the data in the above time periods, the variable is in <month>-<year> format.

Data and Problem Details

Objective: You have to build a data model to predict the probability of default, and choose a cut-off based on what you feel is suitable. Alternatively you can also use a modelling technique which gives binary output.

You have to do the following:

Based on the data that is available during loan application, build a model to predict default in the future. This will help the company in deciding whether or not to pass the loan.

Also note that the data contains defaulters, successful payers and customers who were current during that time. To simplify the problem, customers under 'current' status have been considered as non-defaulters in the dataset.

You will be provided with:

- Dataset containing both train and test data
- Data dictionary