Find attached 3 ﬁles:

**Assessments -** has all the variables created during a loan applicant's risk assessment (i.e. when an approve/reject decision is made)

**Loans -** has data about loans taken by our users

**Data dictionary -** contains a rough description of columns used in both assessments and loans

A user can have multiple assessments - before or after the loan is taken. Only assessments before a loan is disbursed decide whether the loan will be given. The dataset is randomly sampled on both - loans and assessments - so some loans might not have a corresponding assessment before disbursal and similarly, some assessments might not have a corresponding loan.

Loans and assessments are joined by the key 'master\_user\_id'. Every loan has a **max\_dpd** column against it - it is the max of days past due on all instalments of the loan. This is what you can use to deﬁne if a loan has defaulted or not. You can look at 30 days past due as the deﬁnition of default.

The problem statement is to design a model to predict whether a user will default.

As you go through the assignment, I advise noting down the steps/analysis you did, your hypothesis, your ﬁndings etc. Code for model development will be an ideal solution. We want to look at the approach more than the solution.