Description of the Lending Club dataset.

It covers the loan records of the Lending Club from 2007 to the fourth quarter of 2018, which is the largest P2P platform in the US.

Feature selection and the target variable are determined based in studies [An Interpretable Loan Credit Evaluation Method Based on Rule Representation Learner](https://link.springer.com/chapter/10.1007/978-981-99-2385-4_44).

The target variable is loan status, including three values: Current, Fully paid and Charged off. What we actually need here is the record of Fully paid and Charged off. Fully paid refers to the loan has been repaid in full, encoded as 0 in label column, also refers to negative sample in the paper, and Charged off encoded as 1, refers to positive sample, they are applicants of loans that have not been repaid within a period of time. Among 1,345,310 records, we randomly selected 53,812 (4%) of them for the experiment to save training time.

Next, removed percent symbols from Interest Rate and Revolving Utilization Rate columns in the loan records and converted those columns to floats. All the table attributes selected and categories belong are as follows:

Loan Information: Installment, Loan Purpose, Loan Application Type, Interest Rate, Last Payment Amount, Loan Amount, Revolving Balance.

History Information: Delinquency In 2 years, Inquiries In 6 Months, Mortgage Accounts, Grade, Open Accounts, Revolving Utilization Rate, Total Accounts, Fico Range Low, Fico Range High.

Soft Information: Address State, Employment Length, Home Ownership, Verification Status, Annual Income.

Therefor, The data consists of 1,345,310 rows and 22 columns: 21 features and 1 target variable.