Assignment

The objective is to deliver one dataset that can be used to build a credit risk model. Table 1 shows which variables should be included in the dataset and what the datatypes of the variables are. The data you will receive is spread over multiple tables. Note that there are also data quality issues present in the data that need to be solved. Think of missing values, different data formats, different column names, etc. You have to solve these issues and join all the sets together to one final dataset. We would like to ask you to upload your code one day before the presentation.

The different tables contain information on loan payments, demographic factors, credit history, history of the borrowers, and many more variables from loan data provided by one of our clients in the period from 2007Q1 to 2020Q1. The tables contain almost sixty thousand observations; each record with a different loan, representing an unique borrower. Note that all variables contain historical information, including the target variable ("loan_status"). This target variable indicates whether the loan of a borrower is charged off (default) or fully paid.

At the end of the case you are asked to share the final dataset with us and all the steps you have taken to create this set. At the start and midway the case you have the opportunity to ask questions.

Table 1 Variables of final dataset including data types

Variable	Datatype	Description
loan_status	boolean	Current status of the loan (Charged off = DEFAULT (1)
loan_amnt	integer	The listed amount of the loan applied for by the borrower. Charged off (default) or Fully paid (no default)
term	integer	The number of payments on the loan
int_rate	double	Interest rate on the loan
installment	double	The monthly payment owed by the borrower
sub_grade	char	Assigned loan subgrade
emp_length	integer	Employment length in years. Possible values are
home_ownership	char	The home ownership status provided by the borrower.
is_mortgage	boolean	
is_rent	boolean	
is_own	boolean	
is_any	boolean	
is_other	boolean	
annual_inc	integer	The self-reported annual income provided by the borrower.
verification_status	char	Indicates if income was verified or not verified,
is_verified	boolean	

Variable	Datatype	Description
is_not_verified	boolean	
is_source_verified	boolean	
issue_d	date	The month in which the loan was funded to borrower
purpose	char	A category provided by the borrower for the loan request.
addr_state	char	The state provided by the borrower in the loan application
dti	double	A ratio calculated using the borrower's total monthly
fico_range_low	integer	The lower boundary range the borrower's FICO at loan
fico_range_high	integer	The upper boundary range the borrower's FICO at
open_acc	integer	The number of open credit lines in the borrower's credit file.
pub_rec	integer	Number of derogatory public records
revol_bal	integer	Total credit revolving balance
revol_util	double	Revolving line utilization rate, or the amount of credit the
mort_acc	integer	Number of mortgage accounts
pub_rec_bankruptcies	integer	Number of public record bankruptcies
age	integer	The age of the borrower at the time of application
pay_status	integer	Last known repayment status (-2 and -1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)