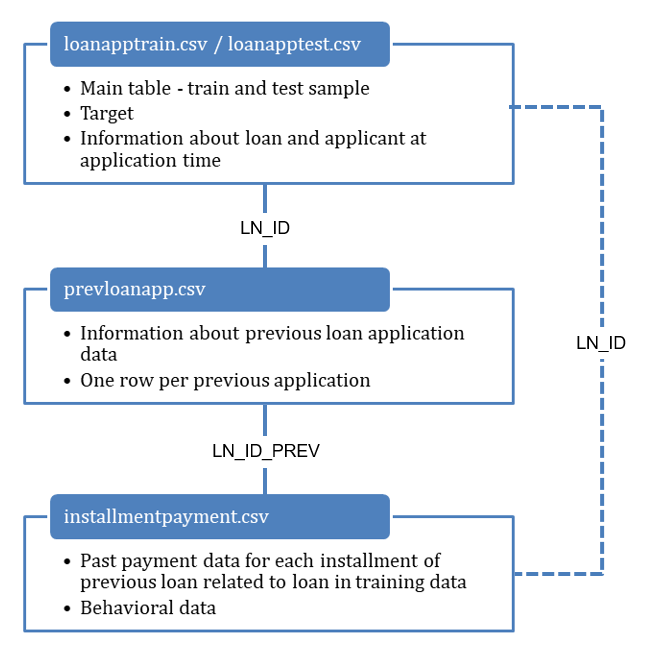
#### (70 point) For most financial institutions, such as banks and multi-finance companies, their main source of income is coming from their lending activities. By engaging in this activity, it means that lenders are exposed to the potential risk, where debtors stop repaying their loans, causing losses to the lenders. To mitigate this loss, lenders are expected to appropriately choose who are qualified for a loan, at what rate, and at what amount.

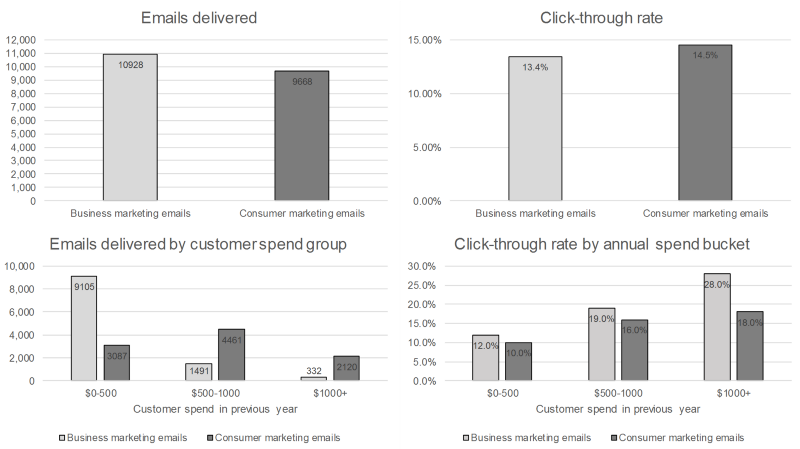
In this question, you are tasked to help the troubled lenders with this problem by creating a model that can help them make their decision. The lenders also provided the following dataset that you can use, which are attached in DS1.zip in the email (columns description is provided in the columns\_description.csv, unnamed columns in the given dataset can be dropped):



Given the dataset above, please create a model to rank the loan applicants based on their repayment capability. The rank will then be used to choose who are eligible for a loan, lower interest rate and higher credit limit. Consider the following points when you are creating the model:

* (10 point) Describe the data pre-processing step that you did
* (5 point) Choose the most appropriate metrics to measure the model performance and provide explanation on why you choose them
* (30 point) Choose 3 of the most important features (original or derived features) and explain how and why they are important
* (15 point) Choose the most appropriate model and provide explanation on why and how the model can solve the lenders problem
* (10 point) Submit the model and all the analysis that you made complete with the test set result (Accuracy, Precision/Recall, F1, AUC, etc)

#### (20 point) A software company is trying to assess their marketing strategy by sending 2 types of marketing emails : business-style email and consumer-style email. We want to see which marketing strategy does better, based on how the emails make user interested in clicking-through. Below are the graphs representing the result of the experiment, The bottom two graphs has the same data as the top two, but binned by the amount of money they spend with the company the year before this emails were sent. Which Campaign did better and why?



1. (10 point) What could be some issues if the distribution of the test data is significantly different than the distribution of the training data?