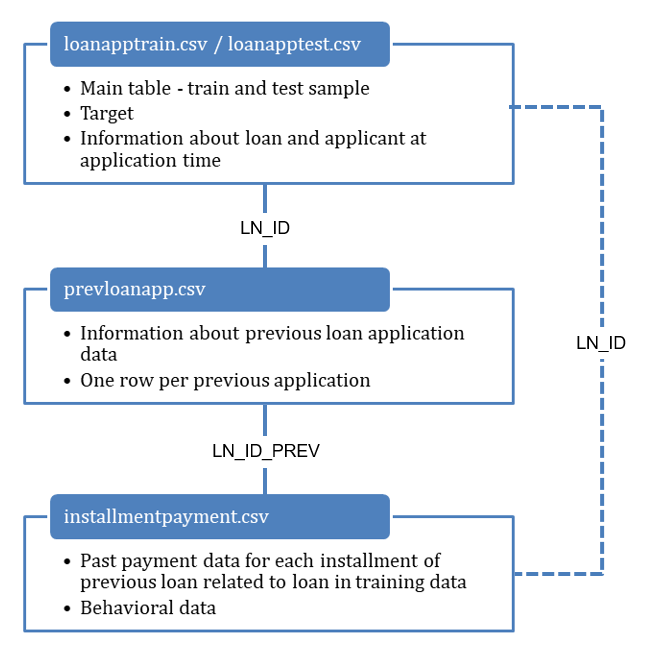
#### (70 point) For most financial institutions, such as banks and multi-finance companies, their main source of income is coming from their lending activities. By engaging in this activity, it means that lenders are exposed to the potential risk, where debtors stop repaying their loans, causing losses 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.

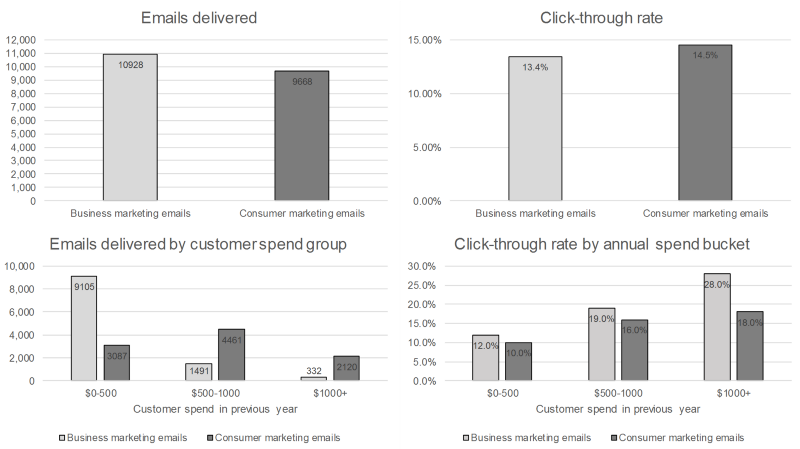
In this question, you are tasked to help the troubled lenders with this problem by creating a model that can help them make their decision. The lenders also provided the following dataset that you can use, which are attached in DS1.zip in the email (columns description is provided in the columns\_description.csv,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

Answer:

1. Data pre-processing
2. Import Library

Pandas : Use for data manipulation and analysis.

Numpy : a fundamental package for scientific computing with python

1. Read Data
2. Data Cleaning

* Missing Value

1. Looking Categorical Value
2. Feature Engineering
3. Correlation Value
4. Machine Learning
5. Important feature
6. Most appropriate metrics

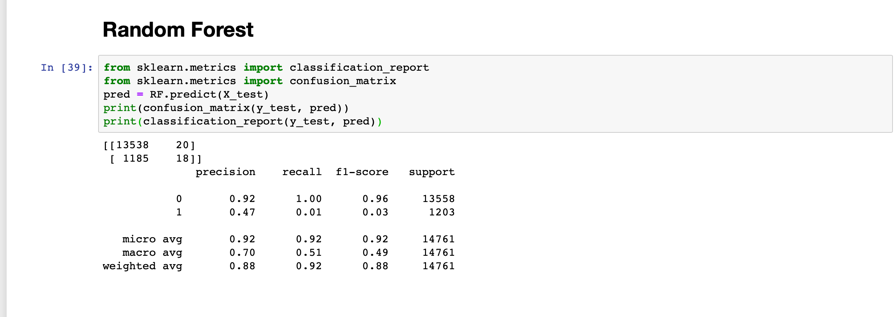
Most Appropriate model to measure = f1-score

We don’t really want to carry both Precision and Recall in our pockets every time we make a model for solving a classification problem. So it’s best if we can get a single score that kind of represents both Precision(P) and Recall(R) which is f1-score.

1. Most Important Feature
2. EXT\_SCORE 1
3. EXT\_SCORE 2
4. PRICE
5. Most appropriate model

Random Forest

1. Test set Result



1. If we want to find which campaign is doing well, then just compare the percentage of click-throughs, and choose the largest percentage of the two data. From the A / B testing results, it is found that the customer marketing emails are doing well.

But from the A / B Testing above the results are still ambiguous, because the total of the two emails sent is out of balance.

Business marketing emails delivered = 10928

Customer marketing emails delivered = 9668

For more accurate A / B Testing, the number of emails sent must be equal, ie equal to ­10928, and also for emails sent must be balanced in 3 categories of customer spend groups. For example, from the range of $ 0-500 the number of Business e-mails sent was 9105 and the number of Customer emails sent were 3087, of course this causes the Testing that we do is not very accurate.

1. If the distribution of the test data is significantly different than the distribution of the training data , “ the prediction will be wrong “.

This significant difference in distribution can occur due to several causes , such as:

1. Sample selection bias

The training examples have been obtained through a biased method, such as non-random split of data into train and test.

If we have a large static dataset, then we should randomly split it into train/test data , and the distribution of test data should be similar to training data

1. Covariate shift

Sometimes the training data and the test data are derived by different processes – eg a food tested on one population is given to a new population that may have significant differences . As a result a classifier based on training data will perform poorly.