Project Using Python

Solving this problem will help you to gain a basic understanding of the credit industry and also how analytics and data science is applied in practice in the BFSI domain.

A short primer of underwriting in the credit industry -

In general, whenever an individual/corporation applies for a loan from a bank (or any loan issuer), their credit history undergoes a rigorous check to ensure whether they are capable enough to pay off the loan (in this industry it is referred to as credit-worthiness).

The issuers have a set of model/s and rule/s in place which take information regarding their current financial standing, previous credit history and some other variables as input and output a metric which gives a measure of the risk that the issuer will potentially take on issuing the loan. The measure is generally in the form of a probability and is a risk that the person will default on their loan (called the probability of default) in the future.

Based on the amount of risk that the issuer is willing to take (plus some other factors) they decide on a cutoff of that score and use it to take a decision regarding whether to pass the loan or not. This is a way of managing credit risk. The whole process collectively is referred to as underwriting.

NOTE : What I have described above is a very simplified version of the process to give you an introductory overview. You can refer to the following resources for a better understanding :

http://budgeting.thenest.com/mean-loan-goes-underwriting-23201.html

http://www.investopedia.com/ (a great source to find meanings of BFSI terminology and jargon)

Overview of the problem

In this project you will have to put yourself in the shoes of a loan issuer and manage credit risk by using the past data and deciding whom to give the loan to in the future. The text files contain complete loan data for all loans issued by XYZ Corp. through 2007-2015. The data contains the indicator of default, payment information, credit history, etc.

The data should be divided into train (June 2007 - May 2015) and out-of-time test (June 2015 - Dec 2015) data. You will have to use the training data to build models/analytical solution and finally apply it to test data to measure the performance and robustness of the models.

You should use the variable : 'issue_d' to divide the data in the above time periods, the variable is in <month>-<year> format.

Data and Problem Details

Objective : You have to build a data model to predict the probability of default, and choose a cut-off based on what you feel is suitable. Alternatively you can also use a modelling technique which gives binary output.

You have to do the following:

Based on the data that is available during loan application, build a model to predict default in the future. This will help the company in deciding whether or not to pass the loan.

Also note that the data contains defaulters, successful payers and customers who were current during that time. To simplify the problem, customers under 'current' status have been considered as non-defaulters in the dataset.

You will be provided with :

- Dataset containing both train and test data
- Data dictionary

Steps to be followed (tentative time required) :

1) Understand the problem and objective (1 hour)
2) Understand the data and develop some business sense. (4-5 hours)
3) EDA, segmentation (if you think is required in this case). (5-6 Hours)
4) Data Cleaning (4-5 Hours)
5) Feature engineering (4-5 Hours)
6) Model Building (try various techniques and at the end justify why you chose a particular technique over others) (3-4 hours)
7) Testing and Cross-validation (3-4 hours)
8) Final results, recommendations and plots/visualizations. (4-5 hours)
9) BONUS : Any other insights or recommendations that you can give from the data which will help the business. (subjective)

Preparing the deck : 6-7 hours

[Actual time might vary from person to person and step to step, this is just indicative]

The final solution should be in the form of a deck showing all the steps above. It will be judged on the following criteria:

1) How well have you adhered to the modelling process discipline.

2) Do your results make business sense, how have you used business intuition to take decisions during the modelling exercise, including but not limited to the following:

- Deciding segmentation (if you choose to have segmentation
- EDA, Feature engineering
- Choosing variables to be put in models
- Deciding a cut-off

3) Performance of your model on test data

- Precision
- Recall
- AUC
- Any other metric you can find from your experience or literature.

4) Estimated monetary impact of your solution (Use the data and your model results and try to come up with a monetary figure that the company can expect to save if they use your model, in business problems it is of a lot of importance since it helps the management take decisions).

Grading system

The final submission should have the following components:

1) All codes (properly documented) - 10%

2) A deck summarizing what they did in each of the above steps.

- Steps 3 to 5 : 30%
- Steps 6 to 7 : 30%
- Step 8 : 30%

3) Bonus marks will be awarded for step 9.