**Exercise: Predict a loan’s choice between Fannie Mae and Freddie Mac**

**Description:**

When a house is being purchased, the borrower applies for a home loan that can eventually get funded by either a GSE[[1]](https://mail.google.com/mail/u/0/" \l "m_-5550454078955576957__ftn1" \o "" \t "_blank) (like FannieMae or FreddieMac) or other financial companies. We are interested in Fannie Mae’s market share and the factors that determine which GSE a loan will go to. The objective of this exercise is to estimate the likelihood that a loan will end up with Fannie Mae, rather than Freddie Mac.  Your model will make the prediction using the features in a public data set called the PUDB.

**Data Portal: The Exchange**

You may get the data with associated data dictionary from this link:

<https://github.com/The-Exchange-Fannie-Mae/National-File-B-Sample-File>

Alternatively, you can use an API from Data Exchange portal to retrieve the data. To do that, first register as a user on <https://developer.theexchange.fanniemae.com/>  and follow the resource guide how to access the API using Python using below link

<https://github.com/The-Exchange-Fannie-Mae/theexchange-python3-clients>

**Data set:**

The data to be used for this exercise is “**Single-Family Public Use Database (PUDB) – National File B**” API  data ( Refer to the API section “ **Loan Data Tool Kit** “) on The Exchange

Note that in the ‘Co-borrower Gender’ field, the category 4 (not applicable) and category 5 (no co-borrower) should be treated as the same category and not used as distinct categories for modeling and projection as ‘4’ is the label reported by Fannie Mae and ‘5’ by Freddie Mac.

**Requirements:**

1. Randomly select 50% of the data as a training set and the remaining 50% as a test set.
2. Use the training set to estimate the probability that a loan is a Fannie Mae loan.

**Work to be delivered:**

1. Python code (with clear notes) and results in a Jupyter notebook
2. Explain the steps and methods you use in data processing, analyses, and modeling
3. The outputs should include
   1. Probability that a loan will be a Fannie Mae loan
   2. Metrics for evaluating model performance on the test set such as confusion matrix, AUROC or other accuracy measures
   3. Charts and tables to illustrate your results if applicable
4. Explain your outcomes and insights obtained from the model. Show which features are most influential to your model prediction