Varo Data Science On-site/Take Home Exercise

The purpose of the exercise is to assess your problem solving and data science modeling expertise. The data challenge is to build a loan default prediction model using publicly available loan data. Please read the instructions below carefully to make your modeling process easier.

Details about the dataset:

- The dataset that will be used for this challenge is the Freddie Mac Single Family Loan-level Dataset, which is publicly accessible on the Freddie Mac website, but is provided you in the zip historical_data_2009Q1.zip. It consists of:
    - historical_data_2009Q1_v3.txt - the "Origination" data file. It contains data for loans originated in the first quarter of 2009, including data about the borrower at the time of origination.
    - historical_data_time_2009Q1_v3.txt - the "Monthly Performance" data file. It contains monthly loan payment data for each of the loans in the Origination data file. You can join the Origination data with the Monthly Performance data via the 'loan_sequence_number'.
- We've also attached Freddie Mac's user guide to the the different columns in the dataset, user_guide_v3.pdf. The key information is on pages 7 onward.

Prompt:

1. Develop a model that predicts whether a loan will go delinquent at any point within 5 years of origination, using data available at the time origination. Describe the model training and feature selection process, and provide an explanation for which model you would recommend. Please assemble your key insights in the form of a Jupyter notebook or presentation that can be sharedwith the Varo team. Please submit your code as part of the completed exercise.

If you use jupyter notebook, please print it as a pdf