# AI in Finance

Professor Hui Chen 15.S52 IAP 2025

## Team Project

**Instructions**

- This project is due by the end of January 24, 2025. Please submit your report and code on Canvas.
- You are encouraged to work in a team of up to 4 students on this project.
- Prepare your report in a format that is concise and easy to read. The report should be typed, double-spaced, with 12 point font, and no more than 4 pages, excluding the cover page and appendix. Additional details (tables, graphs, etc.) can be included in an appendix.

### Project Description

Congratulations! You have been appointed as the Chief Data Scientist at AI-Bank.

You are in charge of developing a fully automated mortgage lending system, which will decide when to approve or reject a mortgage application. The bank hopes that this new system will help maximize the expected profit from its mortgage lending business.

You have access to a dataset of historical information on a large number of single-family 30-year fixed rate mortgages. The dataset includes various loan and borrower characteristics at the time of origination, as well as the loan outcomes (whether it became delinquent in the 3 years after origination). In addition, state-level house prices and national-level interest rates are also included. See "Description.pdf" for details of the datasets and variables.

Potential features to include in your model:

- Loan-level variables: source (FN or FD); quarter of origination; original note rate; original loan-to-value; original combined loan-to-value; original debt-to-income; credit score; mortgage insurance coverage; flag for first-time homebuyer; number of buyers; number of units.
- House prices: past one-year and three-year house price growth at state level.
- Interest rates: current 3-month treasury bill rate; current 30-year fixed mortgage rate; their respective changes from a year ago.

1. Randomly select 70% of the observations from each year to form the training set. Build a logistic model to predict the probability that a loan will become delinquent (at least 30 days past due) within the next 3 years. You can decide how to select the features, but please explain the steps involved.

2. Try to interpret the estimated model. Which variables are important default indicators?

3. Using the remaining 30% of the sample as the test set, produce the confusion matrix first by using decision threshold $\bar{p} = 0.5$, and then $\bar{p} = 0.2$. Discuss the difference that the threshold makes. Discuss how you would choose the optimal $\bar{p}$.

4. (Optional) Redo (a) and (b) using an alternative model of your choice.