NYU

# Mortgage Credit Risk Prediction using Machine Learning Models

Yiming Jia, Albert Kong, Smruti Nalawade, Peter Zhong, Ziyue (Tom) Zhou

March 2025

# Our Project

- **Team Members**
  - Yiming Jia, Albert Kong, Smruti Nalawade, Peter Zhong, Ziyue(Tom) Zhou
- **Background**
  - Traditional ECL models rely on linear assumptions, limiting their adaptability to complex borrower and market dynamics.
- **Opportunity**
  - Machine learning techniques can integrate broader data (macroeconomic, geospatial) to enhance accuracy and insight. Using machine learning models, we can make better predictions for mortgage credit risk.

# Project Objectives

- **Loan Default Prediction**
  - Improve Probability of Default (PD) accuracy through advanced ML.
- **Credit Loss Prediction**
  - Refine ECL estimates by leveraging actual loss data from Freddie Mac.
- **Delinquency Forecasting**
  - Integrate macroeconomic indicators to anticipate delinquency trends.
- **Geographic Risk Analysis**
  - Identify and visualize high-risk regions using geospatial modeling.

# Dataset Overview

- **Freddie Mac Single-Family Loan-Level Data**
  - Origination details (credit scores, LTV, DTI)
  - Monthly performance metrics (delinquency status, foreclosure events)
  - Actual loss information for defaulted loans
- **St. Louis Fed (FRED) Macroeconomic Data**
  - GDP, unemployment, interest rates, inflation
  - Aligned by time periods (monthly/quarterly)
- **Data Integration**
  - Merged on date fields to correlate loan performance with economic indicators

# Methodology

- **ECL Computation**
  - ECL = PD × Predicted Actual Loss
  - Uses model-estimated PD + Freddie Mac actual loss data
- **PD Estimation**
  - Classification algorithms (e.g., Random Forest, XGBoost)
  - Key features: borrower characteristics, loan terms, macroeconomic variables
- **Delinquency Forecasting**
  - Modeling delinquency trends as a high-dimensional time-series forecasting problem to forecast potential risks and market shifts by leveraging statistical ML models (ARIMA, VAR) or deep learning models (LSTM, transformers, etc).
  - Combines historical performance + macroeconomic factors
- **Geospatial Risk Insights**
  - Clustering methods (K-Means, DBSCAN) to detect high-default regions
  - Heatmaps to visualize geographic concentrations of risk
  - State-level or county-level breakdowns to inform local strategies

# Tech Stack / Tools

| Category | Tools & Technologies |
|---|---|
| **Data Processing & Management** | Pandas, NumPy |
| | SQL (SQLite) - if needed |
| **Machine Learning & Modeling** | Scikit-Learn |
| | XGBoost, Random Forest |
| **Time-Series Forecasting** | Statsmodels, Survival Analysis |
| **Geospatial Analysis & Visualization** | GeoPandas, Folium |
| | Seaborn, Matplotlib |
| **Project Management & Collaboration** | Agile Methodology (Scrum/Kanban) |
| | GitHub |
| | Jupyter Notebooks, VS Code |
| **Documentation & Reporting** | Google Docs, Google Sheets |
| | Microsoft PowerPoint (PPT) |

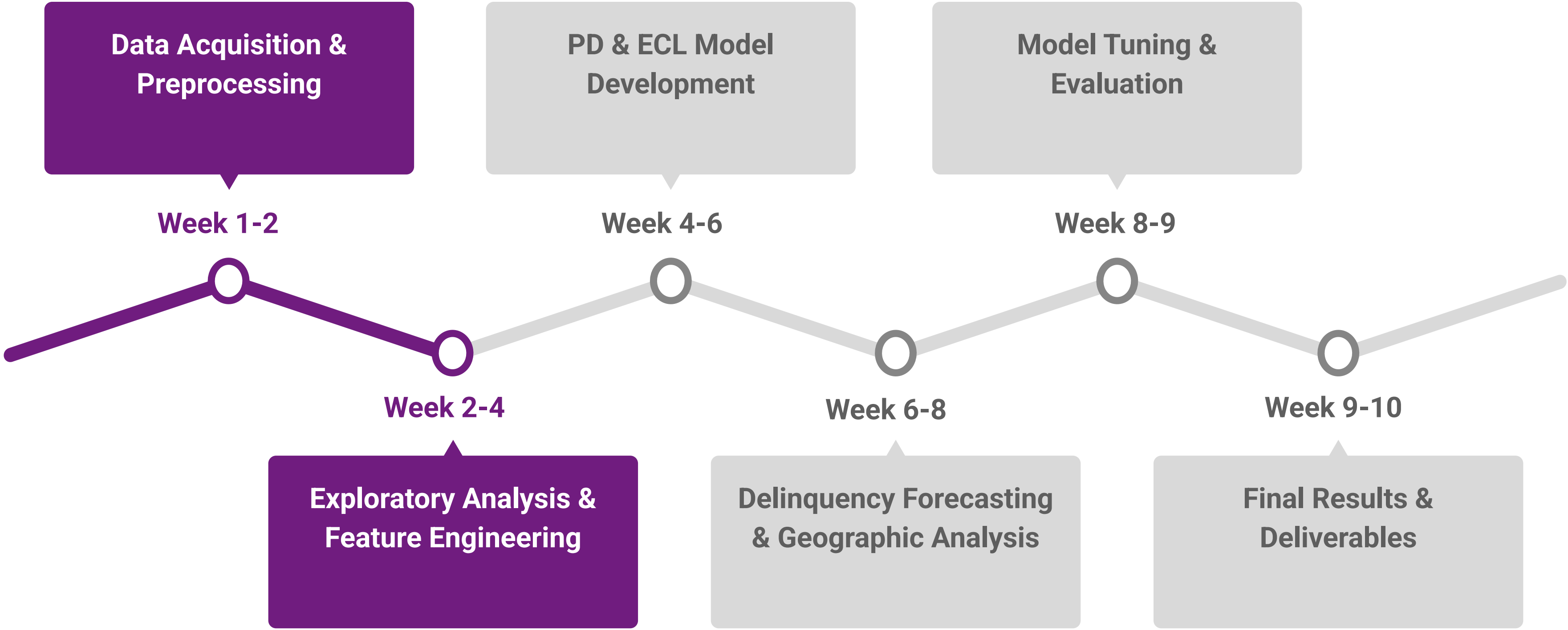
# Python Optimization Techniques

- **Performance Profiling**
  - cProfile, memory_profiler to identify and address bottlenecks
- **JIT Compilation**
  - Numba for faster loops and numerical computations
- **Parallelization**
  - Multi-threading (I/O tasks) and multi-processing (CPU tasks)
- **Vectorization**
  - NumPy/Pandas for bulk operations on large datasets
- **GPU Acceleration**
  - CuPy for heavy matrix operations, boosting speed

# Outcomes

- **Improved Default Prediction**
  - Higher accuracy in PD estimation benefiting financial institutions’ underwriting decisions and researchers’ analyses
- **Refined ECL Estimates**
  - More precise loss forecasts for proactive provisioning
- **Risk Insights**
  - Early detection of delinquency trends, identification of geographic “hot spots”
- **Efficiency & Scalability**
  - Advanced parallelization and GPU acceleration for handling massive loan datasets

# Project Timeline



Week 1-2

Data Acquisition & Preprocessing

Week 2-4

Exploratory Analysis & Feature Engineering

Week 4-6

PD & ECL Model Development

Week 6-8

Delinquency Forecasting & Geographic Analysis

Week 8-9

Model Tuning & Evaluation

Week 9-10

Final Results & Deliverables

# Any Questions?

## Thank You!