Lecture 1

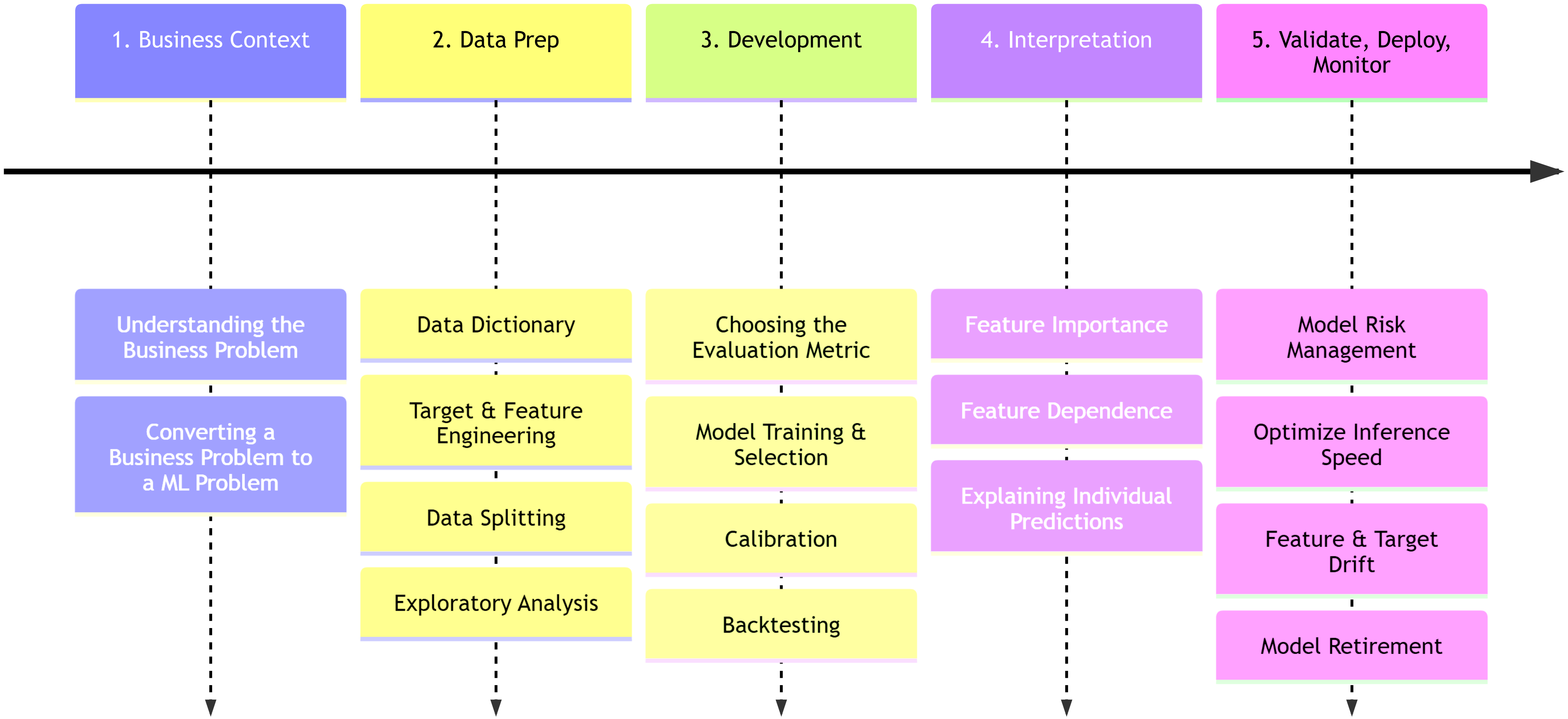
Interpretable Machine Learning for Finance

Instructor

William Chiu, FRM

- UC Berkeley Econ Alum (Class of 2005)
- Financial Risk Manager (FRM) Certified since 2012
- Northwestern MLDS Alum (Class of 2013)
- Senior Credit Risk Analyst at a credit bureau (2013-2014)
- Head of Statistical Modeling at a bank (2014-2023)
- VP of Predictive Modeling at an insurance company since 2023

Model Lifecycle



1. Business Context

Business Problem	ML Problem
Acquire new clients through a direct-mail campaign	Develop a marketing response model and target high-response prospects
Approve or reject loan applications based on risk	Develop a credit scorecard model and approve low-risk applications
Explain why a loan application was rejected	List the differences between an approved application and a rejected application that are otherwise very similar (i.e., a counterfactual)
Block fraudulent transactions	Develop a cost-sensitive model and block high-risk transactions

2. Data Prep

1. Borrower data from an Indian Bank and Credit Bureau. The instructor modified the data to include `campaign_id`, `control_group`, and `response_flag`. (Azad 2024).
2. Loan Performance data from LendingClub. Includes both accepted and rejected loan applications. (George 2019)
3. Simulated data from the *Fraud Detection Handbook* . (Le Borgne et al. 2022)

3. Model Development

- AutoGluon is an open-source AutoML framework designed to automate model training and ensembling.
- AutoGluon achieves strong performance by using ensembling techniques and a comprehensive library of preset hyperparameters for each base learner.
- AutoGluon offers data preprocessing like automatic encoding of categorical variables and automatic removal of features with no variance.
- Scikit-learn provides calibration methods (e.g., Platt scaling, isotonic regression) to improve the reliability of predicted probabilities.

4. Interpretable ML Tools

- Rank features from most to least important
  - Permutation importance
  - Aggregated SHAP importance
  - Dealing with multicollinear features
- Describe the dependence between features and predictions
  - Partial dependence plots (PDP)
  - Accumulated local effects (ALE) plots
  - SHAP dependence plots

4. (More) Interpretable ML tools

- Explain individual predictions (local explanations)
  - SHAP values
  - Counterfactuals

Prerequisites

- Comfort using supervised learning methods:
  - Generalized Linear Models
  - Trees with bagging and/or boosting
  - Cross Validation and Bias-Variance Tradeoff
- Familiarity with Python and data science libraries (i.e., pandas and scikit-learn)
- Practice with SQL databases

Assignments and Grading

1. Problem Set 1 (25%)
2. Problem Set 2 (25%)
3. Problem Set 3 (25%)
4. Problem Set 4 (25%)

Introduction to Retail Banking

What is a Retail Bank?

- A retail bank serves individual consumers, households, and small businesses
- Offers deposit and lending products: checking/savings accounts, certificates of deposit, credit cards, and mortgages
- Large retail banks also offer brokerage and wealth management services

Revenue Sources

- Interest collected from loans (e.g., credit cards, auto loans, personal loans, mortgages)
- Transaction fees, commissions, and service charges
- Major contributor to profitability: spread between interest collected from loans and interest paid to depositors

Expense Sources

- Employee compensation
- Interest paid to depositors
- Provision for loan losses

How could a Bank fail?

- Insufficient liquidity (not enough cash to meet obligations)
- Insufficient capital (not enough equity to absorb losses)
- High levels of non-performing loans or significant credit losses
- Sudden loss of depositor confidence leading to a bank run

What is FDIC insurance?

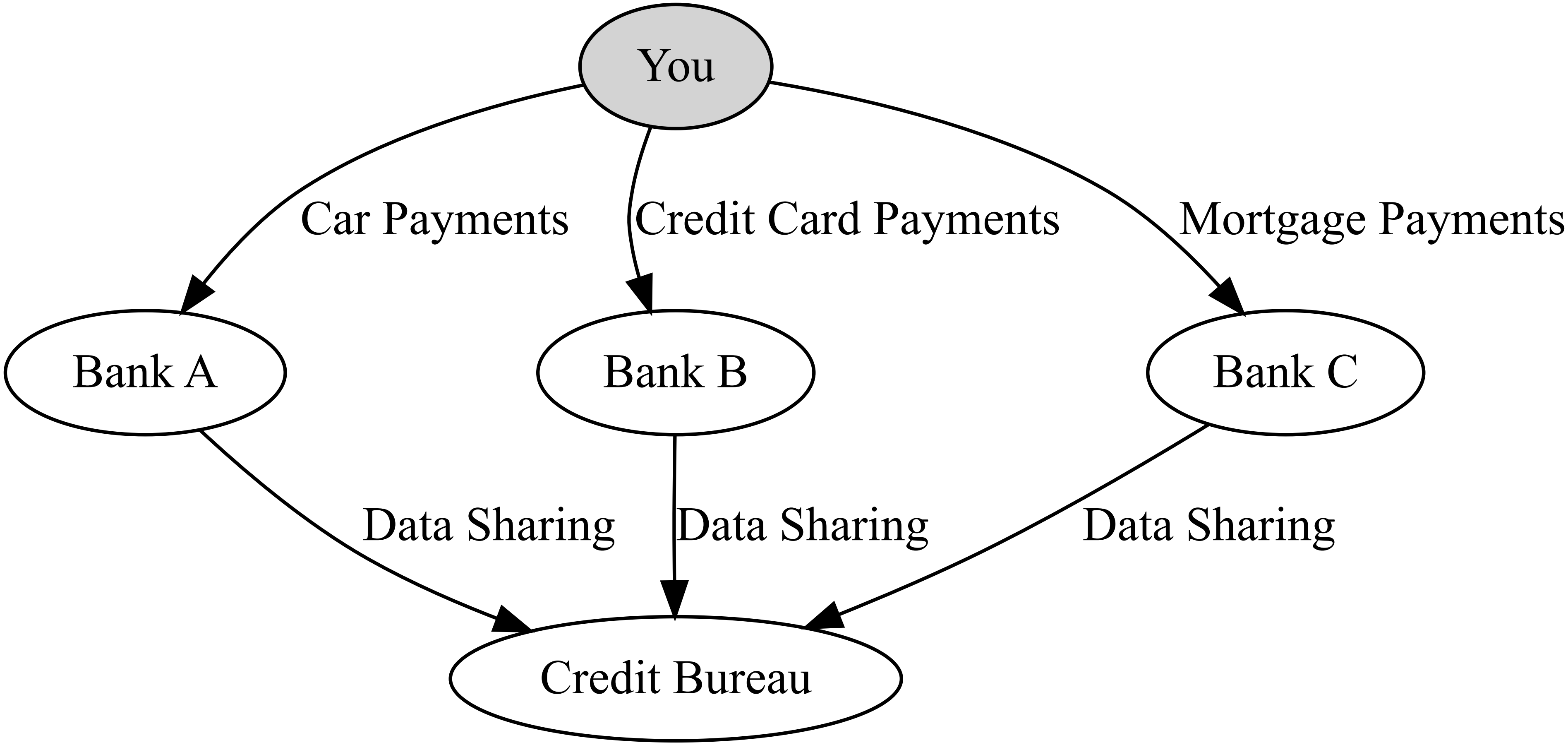
- Protects depositors by guaranteeing their funds up to a certain limit if a bank fails
- Aims to maintain public confidence and prevent bank runs
- Standard insurance amount in the US: $250,000 per depositor, per insured bank

How do Banks use machine learning?

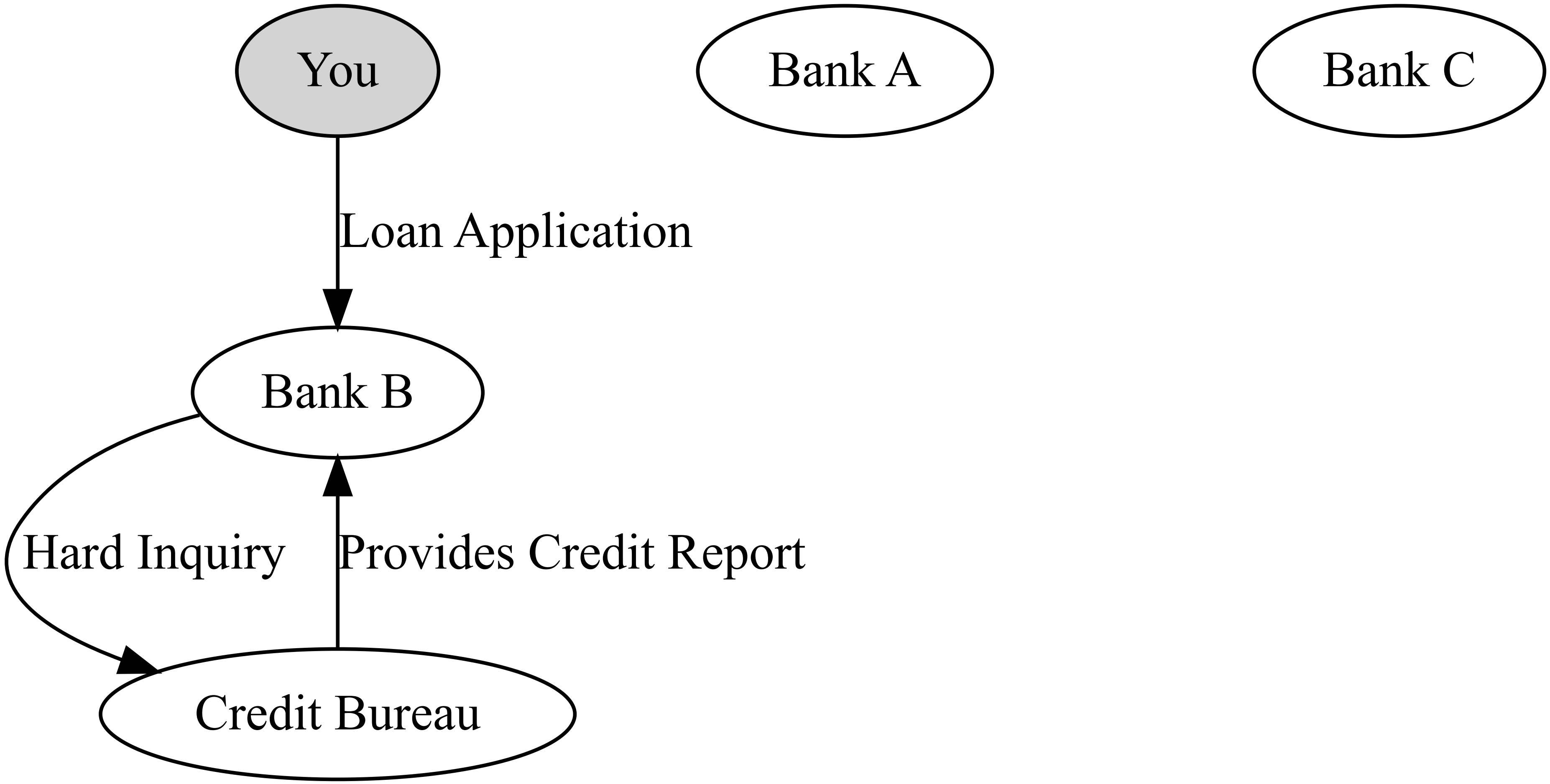
- Improve the effectiveness of marketing campaigns
- Measure the credit risk of loan applicants
- Estimate the lifetime credit losses of a loan portfolio
- Detect fraudulent activities

Introduction to Credit Bureaus

When You Pay Your Loans



When You Apply for a Loan



US Credit Bureaus

- The US has three major credit bureaus:
  - Experian
  - TransUnion
  - Equifax
- Credit bureaus serve three major functions:
  - Collect and maintain consumer credit information
  - Provide credit reports to lenders
  - Generate credit scores

Bureau Data Structure

- **Consumer Information**:
  - Borrower details: Name, address, SSN, birthdate
  - Employment data (if reported)
  - Credit scores from different scoring models
- **Trade Lines (Accounts)**:
  - Account details: Creditor, type, opened/closed dates
  - Credit terms: Limit, payment amount, loan amount
  - Current status: Balance, utilization, payment status
  - Payment history: 24-84 months of payment records

Credit Report Inquiries

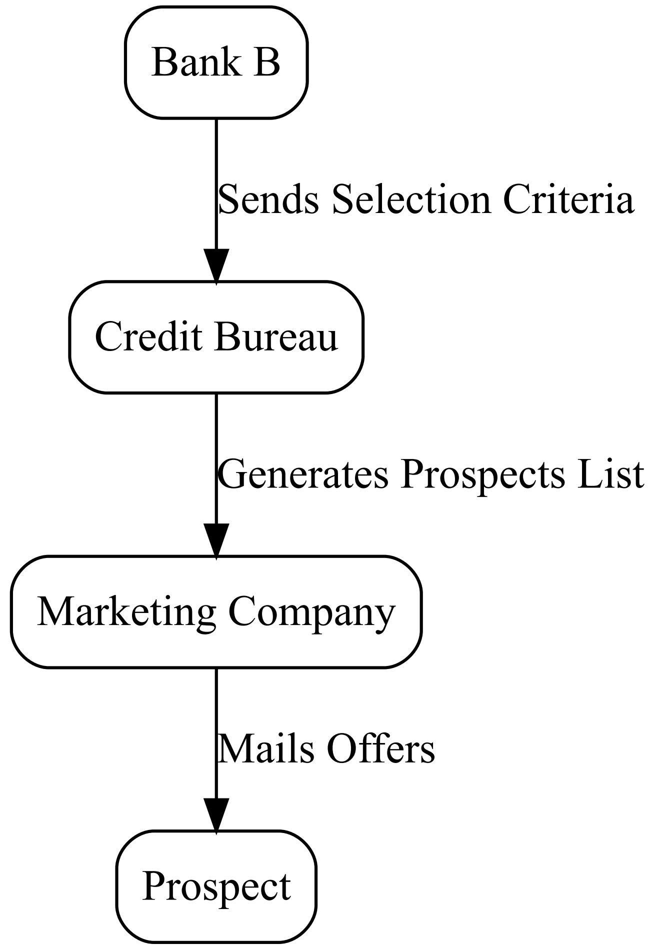
- **Hard Inquiries**:
  - Result from credit applications
  - Visible to all lenders
  - Impact credit scores
- **Soft Inquiries**:
  - Account reviews, promotional screening
  - Only visible to consumer
  - No impact on credit scores

Credit Scores

- Two major scoring models:
  - FICO
  - VantageScore
- Credit Score Ranges:
  - Excellent: 800+
  - Very Good: 740-799
  - Good: 670-739
  - Fair: 580-669
  - Poor: Below 580

Prescreen Campaigns to Acquire New Clients

Prescreen Process



Business Problem

- The marketing team wants to acquire new credit card customers through a direct mail campaign
- The campaign features a **zero-interest rate balance transfer promotion** for 12 months
- Campaign costs include: mail costs, credit bureau fees, and expected credit losses
- The marketing budget is limited, so the marketing team asks you to identify high-response prospects
- Compliance with regulatory requirements (FCRA, ECOA) is necessary

Business Solution

- Develop selection criteria for credit bureau prescreening
  - Credit Score >= 700
  - No delinquent loan payments in the last 60 months
  - At least one active credit card with 60 months of on-time payments
  - Must live within a list of 20 metropolitan statistical areas (MSA)
- In addition to the selection criteria, your model would identify high-response prospects

Machine Learning Problem

- Train a classification model to identify prospects most likely to accept the offer
- Requires historical campaign data, response flags, and relevant bureau attributes
- Develop a model that maximizes AUC and estimate a probability threshold that maximizes the F1 score, precision, or recall
- Score prospects, select high-scoring individuals for the campaign
- Measure the effectiveness of the campaign using a control group

Features to Consider

- Credit Utilization: Higher utilization → Higher response rates
- Revolving Balances: Higher balances → Higher response rates
- Balance Transfer History: Previous transfers → Higher response rates
- Recent Inquiries: More inquiries → Higher response rates

Prescreening Compliance Framework

- Fair Credit Reporting Act (FCRA)
  - Do not prescreen consumers who are on the opt-out list
  - Applicants who meet the prescreening criteria should be approved for credit
  - If the application is declined, provide clear reasons
- Equal Credit Opportunity Act (ECOA)
  - Prohibits discrimination based on race, color, religion, national origin, sex, marital status, or age
  - Apply consistent underwriting standards to all applicants
  - Maintain documentation to demonstrate non-discriminatory practices and facilitate regulatory reviews

Citations

Azad, Rohan. 2024. “Leading Indian Bank & CIBIL Real-World Dataset.” https://www.kaggle.com/datasets/saurabhbadole/leading-indian-bank-and-cibil-real-world-dataset/data.

George, Nate. 2019. “All Lending Club Loan Data.” https://www.kaggle.com/datasets/wordsforthewise/lending-club.

Le Borgne, Yann-Aël, Wissam Siblini, Bertrand Lebichot, and Gianluca Bontempi. 2022. *Reproducible Machine Learning for Credit Card Fraud Detection - Practical Handbook*. Université Libre de Bruxelles. https://github.com/Fraud-Detection-Handbook/fraud-detection-handbook.