

Lecture 1

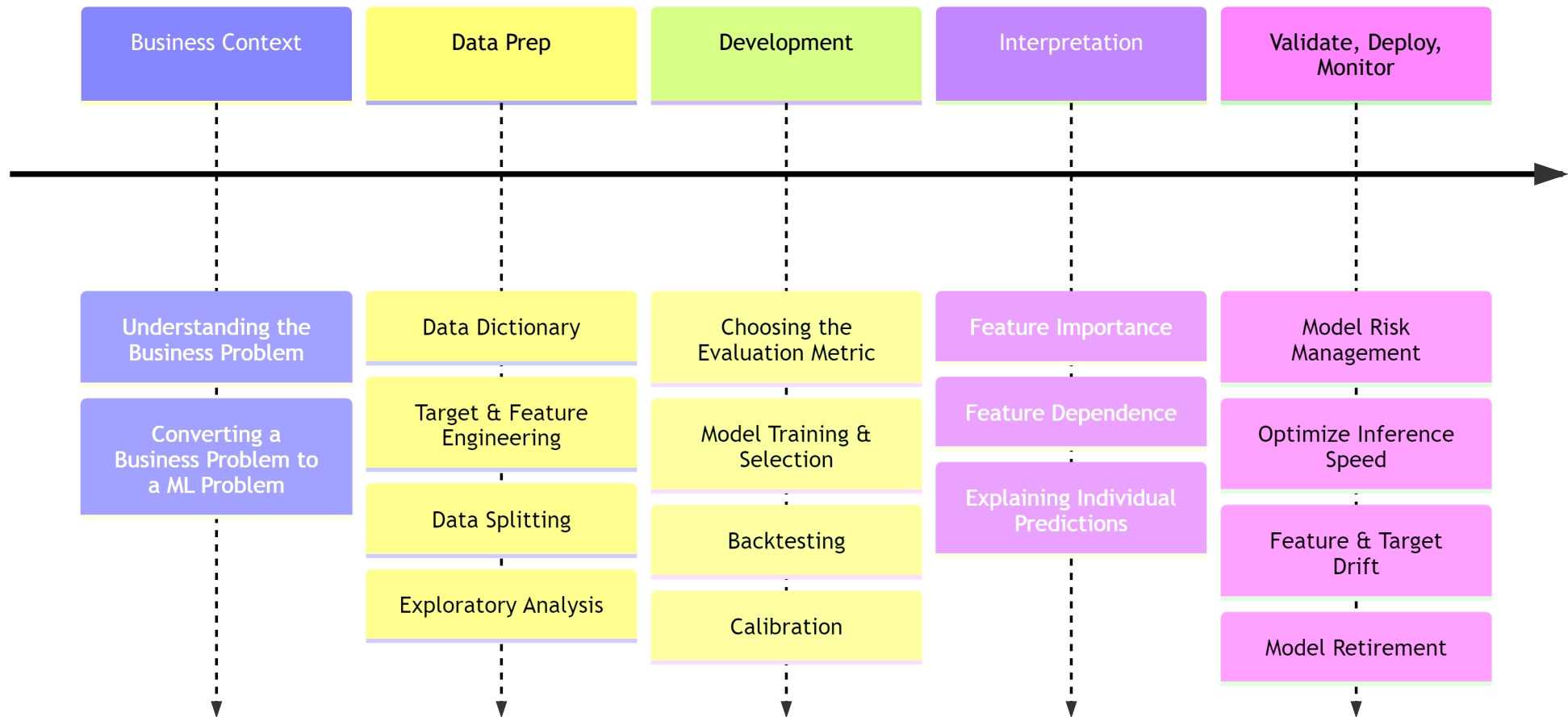
Interpretable Machine Learning for Finance

Instructor

William Chiu, FRM

- Financial Risk Manager (FRM) Certified since 2012
- MLDS Alum since 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



Business Context

1. Acquire new clients with a marketing response model (aka prescreen model)
2. Approve (or reject) loan applications with a credit scorecard model
3. Explain rejected loan applications with counterfactuals
4. Estimate lifetime credit losses with a transition matrix model
5. Detect fraud with a cost-sensitive model



Data Sets

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. Single-Family Loan Performance data from Fannie Mae. ([Fannie Mae 2024](#))
4. Simulated data from the *Fraud Detection Handbook* . ([Le Borgne et al. 2022](#))

Model Development Tools

- AutoGluon is an open-source AutoML framework that automates model training and ensembling.
- Rather than focussing on hyperparameter optimization (HPO), AutoGluon focuses on delivering strong performance through ensembling.

Interpretable ML Tools



- Rank features from most to least important
 - Permutation importance
 - 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

(More) Interpretable ML tools

- Explain individual predictions
 - 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 Sets (70%)

- Weekly conceptual questions and programming exercises based on each week's topics.
- Reinforce understanding of weekly material and practical application of concepts.

2. Kaggle Competition (30%)

- Develop a scorecard model for loan underwriting.
- Apply modeling techniques in a competitive scenario.
- Students will be given a labeled training set and an unlabeled test set.
- Students will be asked why the model rejected a loan application.
- Grades will depend on model performance in an unlabeled test set and the quality of the explanation.

Introduction to Retail Banking

What is a Retail Bank?

- Serves individual consumers and small businesses
- Offers checking accounts, savings accounts, personal loans, mortgages, and other financial products
- Focuses on everyday banking services rather than large-scale corporate or investment activities

Revenue Sources

- Collects interest on loans (e.g., mortgages, auto loans)
- Earns revenue from transaction fees, service charges, and other financial products (e.g., credit cards)
- Major contributor to profitability: spread between interest charged on loans and interest paid on deposits

Expense Sources

- Employee salaries
- Branch operations
- Technology infrastructure
- Regulatory compliance costs
- Capital for loan loss provisions
- Marketing and customer acquisition costs

How could a Bank fail?

- Insufficient liquidity
- Massive loan defaults
- Sudden loss of depositor confidence leading to a bank run
- Poor risk management, reckless lending, or major market downturns
- Regulatory sanctions and reputational damage

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 credit risk assessment
- Detect fraudulent activities
- Optimize marketing campaigns
- Predict customer behaviors, default risks, and repayment probabilities
- Automate processes such as loan underwriting and customer service

Introduction to Credit Bureaus

US Credit Bureaus

- Experian
- TransUnion
- Equifax

Bureau Data

- **Borrowers:** Consumer demographics, employment information, and credit history
- **Trades:** Details on lines of credit, outstanding balances, and payment patterns
- **Trade History:** Monthly updates, delinquencies, and overall account performance

Prescreen Campaigns to Acquire New Clients

Business Problem

- Identify potential borrowers likely to respond positively to credit offers
- Target specific consumers based on attributes like credit profile and spending behavior
- Maximize response while minimizing costs and regulatory risks

Business Solution

- Use targeted mailing lists or digital ads to market credit products
- Selection criteria: credit scores, income levels, past responsiveness to similar campaigns
- Tailor offers to the right audience to grow customer base efficiently

Machine Learning Problem

- Predict which prospects are most likely to accept the offer
- Requires historical campaign data, response flags, and relevant borrower attributes
- Build a model that balances precision (fewer false positives) with recall (capturing most likely responders)

Machine Learning Solution

- Train a classification model (e.g., gradient boosting or logistic regression) on labeled data to identify strong responders
- Incorporate performance metrics such as AUC or recall to gauge effectiveness
- Score new prospects, select high-scoring individuals for the campaign, and measure the uplift over a control group

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>.

Fannie Mae. 2024. “Single-Family Loan Performance Data.”

<https://capitalmarkets.fanniemae.com/credit-risk-transfer/single-family-credit-risk-transfer/fannie-mae-single-family-loan-performance-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>.