EARLY PAYMENT DEFAULT PREDICTION IN MORTGAGE LENDING

PRESENTATION

Capstone Project –
Sections 5–6: Modeling and Results,
Limitations Encountered

Presented by Siobhan Curran

Faculty of Computing and Data Sciences, Boston University

PROJECT OVERVIEW

This project focuses on predicting early payment defaults (EPDs) in mortgage lending.

Three Kaggle datasets were used to compare predictive models and identify borrower segments linked to delinquency risk.

The analysis combined:
- Supervised learning for default prediction (KNN, Gradient Boosting)
- Unsupervised learning for borrower segmentation (K-Means, DBSCAN, HAC)

Finally, I will highlight key challenges I encountered during the report development process.

Table 1 - Summary of Datasets Used for Mortgage Default Analysis

Dataset Name	Source	Approx. Records	Variables	Default Rate	Primary Analytical Use
Home Credit Default Risk	Kaggle (n.d.-b)	307,000	122	7 %	Gradient Boosting model training
Give Me Some Credit	Kaggle (n.d.-a)	150,000	11	8 %	Baseline KNN model testing
Lending Club Loan Data	Kaggle (n.d.-c)	2,200,000	145	25 %	Unsupervised clustering and segmentation

Note. All datasets are publicly available through Kaggle.

Record counts are approximate after data cleaning.

Default rate represents the proportion of positive-class (default) observations in each dataset.

METHODOLOGY

Hybrid Analytical Approach:

- Combined supervised and unsupervised models
- Addressed imbalanced financial data and improved interpretability

Supervised Models:

- K-Nearest Neighbors (KNN): Baseline, distance-based classifier
- Gradient Boosting (GB): Ensemble model capturing complex borrower relationships

Unsupervised Models:

- K-Means, DBSCAN, HAC: Identified borrower segments and outliers
- Key Goal: Improve default prediction and reveal borrower risk patterns

Key Goal:
Improve default prediction and reveal borrower risk patterns

Section 5: Modeling and Analysis

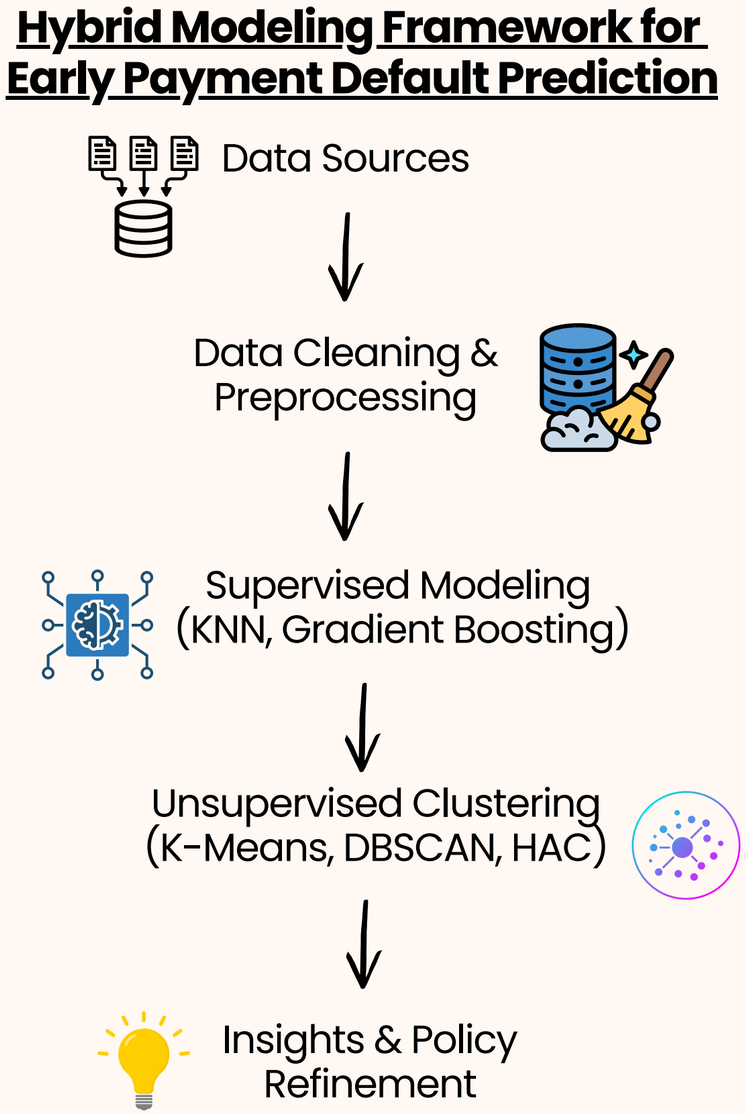

Figure 2. Hybrid Modeling Framework for Early Payment Default Prediction.

MODEL PERFORMANCE & EVALUATION

Key Observations:

- Gradient Boosting (GB) outperformed KNN across all metrics.
- GB achieved **PR AUC = 0.226** and **ROC AUC = 0.740**, surpassing the 0.20 success benchmark.
- Decision-threshold optimization (cutoff = 0.1388) improved F1 score from 0.017 to 0.294.
- KNN achieved high accuracy but poor recall, confirming weak performance in rare-event detection.

Model	Recall	Precision	F1	PR AUC	ROC AUC
KNN	0.09	0.16	0.16	0.2	0.7
Gradient Boosting	0.27	0.31	0.29	0.226	0.74

Table 2. Comparative model performance on imbalance mortgage datasets

Metric definitions: PR AUC = Precision-Recall Area Under Curve;
ROC AUC = Receiver Operating Characteristic.

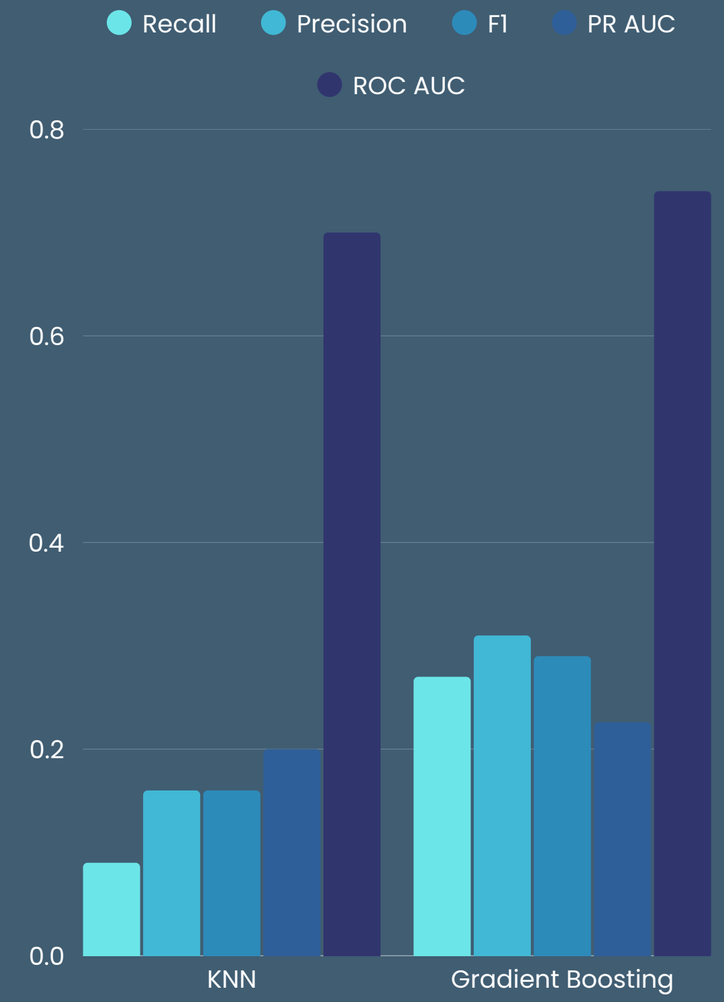

Figure 3. Model performance comparison (KNN vs. Gradient Boosting).

FEATURE IMPORTANCE

Rank	Feature	Description
1	EXT_SOURCE_2	External credit score (risk indicator)
2	EXT_SOURCE_3	*Additional credit score measure*
3	*DAYS_BIRTH*	*Borrower age (in days, negative coded)*
4	*CREDIT_TO_INCOME*	*Credit-to-income ratio*

Figure 4. Top predictive features from the Gradient Boosting model.

Key Insights:

- External credit scores (EXT_SOURCE_2 and EXT_SOURCE_3) were the strongest predictors of default.
- Younger borrowers and those with higher credit-to-income ratios showed higher default risk.
- These results align with industry expectations: credit stability and repayment capacity are the most reliable early indicators of delinquency.
- Feature interpretability supports practical use in underwriting and risk monitoring.

BORROWER SEGMENTATION: UNSUPERVISED LEARNING

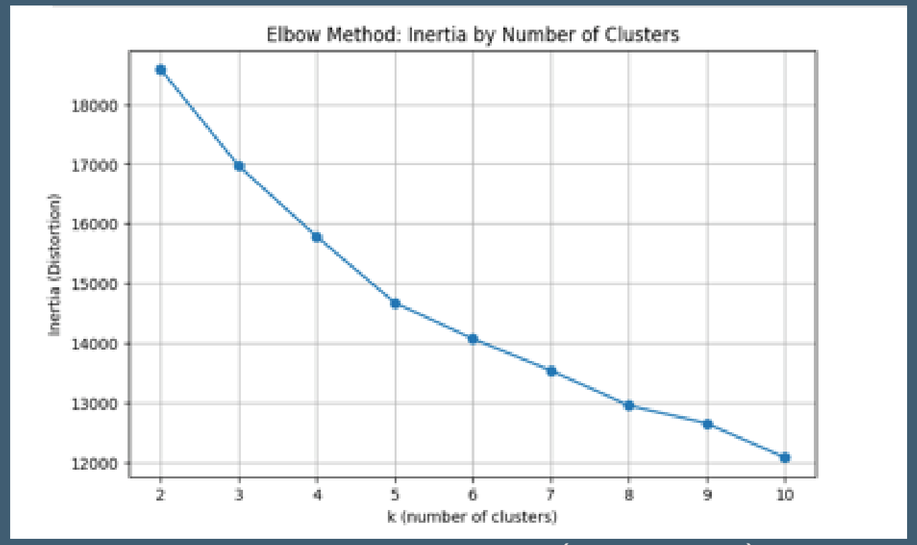

Figure 5. Optimal cluster selection (Elbow Method).

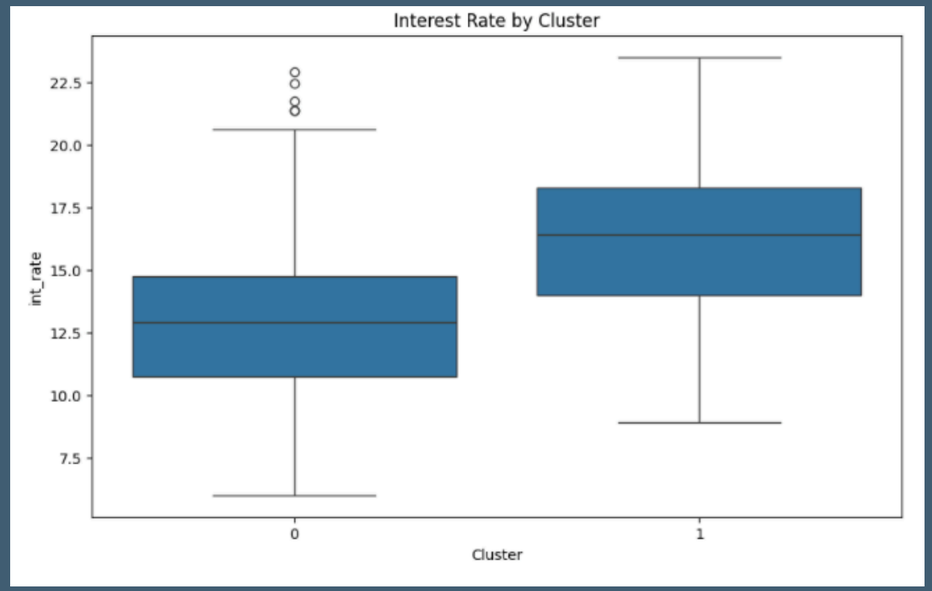

Figure 6. Interest rate variation across borrower clusters.

Key Findings:

- K-Means clustering revealed two main borrower segments (k = 2).
- Cluster 1 borrowers had higher loan amounts and higher interest rates, identifying them as higher-risk.
- DBSCAN confirmed data cohesion and flagged 19 outliers, showing a small population of unique risk cases.
- Hierarchical Agglomerative Clustering validated the same two-cluster structure, confirming model stability.

Interpretation:

These clusters represent meaningful borrower profiles that lenders can use to refine loan approval criteria, set differentiated rates, and strengthen risk monitoring.

BUSINESS RECOMMENDATIONS & POLICY IMPLICATIONS

Key Recommendations:

1. Deploy Gradient Boosting model with the optimized decision threshold of 0.1388 to improve early default detection.
2. Integrate cluster segmentation (Cluster 1) into lending policies to tailor risk controls, loan terms, and pricing.
3. Apply continuous monitoring to maintain model calibration and fairness, avoiding bias in lending outcomes.

Strategic Impact:

1. Improves risk sensitivity and reduces early portfolio losses.
2. Enables data-driven loan approval and pricing decisions.
3. Strengthens compliance through transparent, explainable model design.

From Model Insight to Business Action:

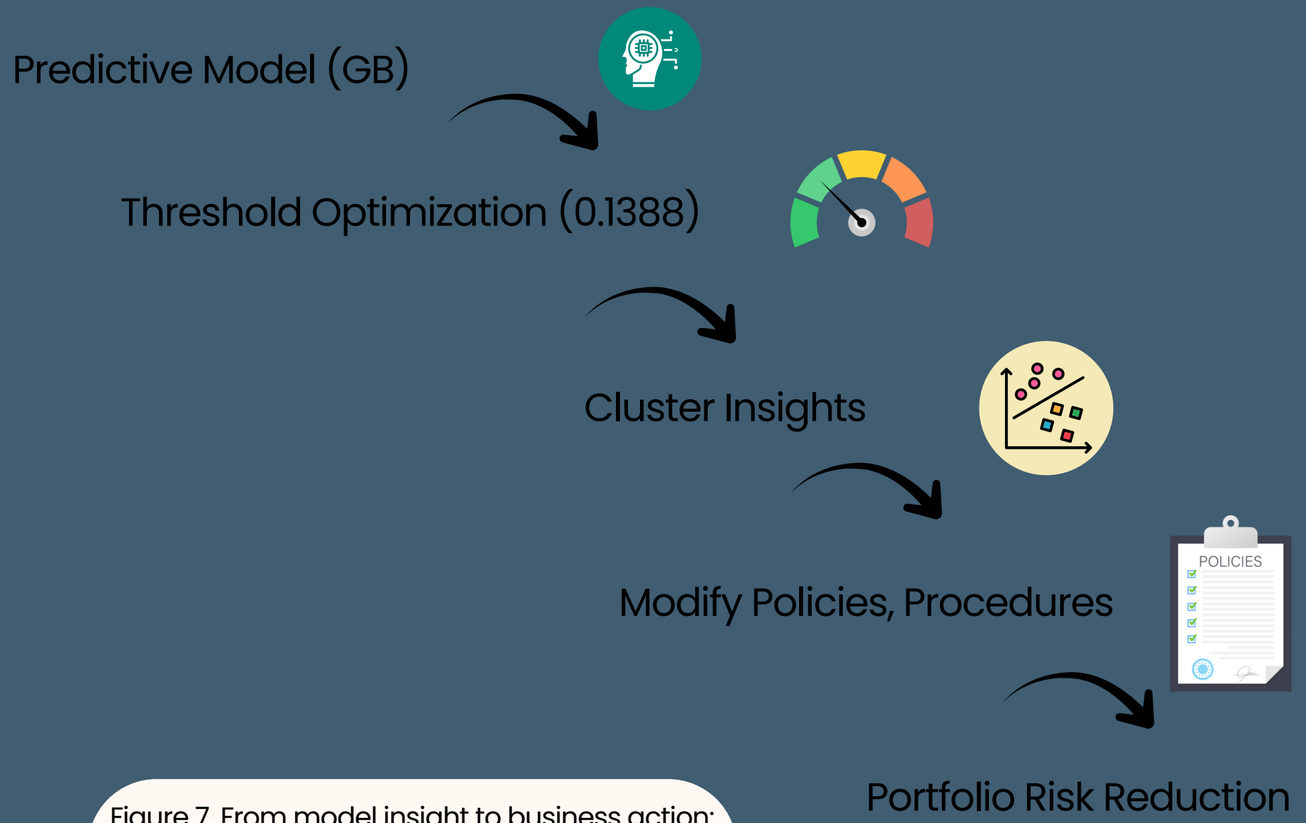

Figure 7. From model insight to business action: linking predictive modeling, threshold optimization, and borrower segmentation to policy refinement and portfolio risk reduction.

CONCLUSION & NEXT STEPS

Key Takeaways:

- The optimized Gradient Boosting model proved most effective for predicting early payment defaults.
- Decision-threshold tuning was critical, improving the F1 score from 0.017 to 0.294, showing measurable gains in sensitivity.
- Unsupervised clustering provided interpretable borrower profiles that can directly inform policy decisions.
- This combined analytical approach improves early detection, supports fair lending, and strengthens portfolio risk management.

Next Steps:

- Deploy the Gradient Boosting model with threshold monitoring.
- Implement segmentation logic for borrower profiling.
- Conduct fairness and calibration audits quarterly.
- Extend data sources to include behavioral or external credit data for further accuracy gains.

LIMITATIONS & PROBLEMS ENCOUNTERED

BALANCING TECHNICAL DEPTH WITH READABILITY

Deciding how much modeling detail to include for both technical and general audiences.

CONDENSING COMPLEX SECTIONS

Reducing modeling and results discussions while keeping clarity and flow within page limits.

MAINTAINING APA AND VISUAL CONSISTENCY

Ensuring formatting, figure captions, and citations were accurate and uniform throughout the document.

THANK YOU

Siobhan Curran

smcurran@bu.edu

Faculty of Computing and Data Sciences, Boston University