Great Learning Capstone Project

LOAN DEFAULT PREDICTION

ENHANCING RISK MANAGEMENT THROUGH DATA-DRIVEN INSIGHTS

Presented By
Natasha Soldin

April 2025

KEY TAKEAWAYS

PROBLEM

Home equity loans show a 20% default rate, affecting 1,192 out of 5,960 loans. With an average loan amount of $18,608, this represents $22.2M in at-risk value annually, directly impacting profitability and capital reserves.

SOLUTION

A dual-model prediction system that combines a Random Forest model for balanced accuracy with a Decision Tree model for interpretability. This is implemented through a tiered approval framework that tailors verification requirements to risk levels.

RESULTS

91% overall prediction accuracy with strong performance across both defaulting and non-defaulting loans. Identification of the strongest predictors: credit history quality, debt-to-income thresholds, and occupation-purpose combinations.

IMPACT

With a conservative estimate of 20% default reduction, the bank can save approximately $4.4M annually. Additional benefits include faster processing times, improved regulatory compliance, and better resource allocation for loan officers.

PROBLEM DEFINITION

01 Financial Risk Exposure

Non-performing loans directly impact profitability through write-offs and provisioning requirements. With $22.2 million at risk annually, even small improvements in default prediction can deliver substantial savings.

02 Process Inefficiencies

Manual, document-heavy approvals create processing bottlenecks and inconsistent risk assessments. This extends approval timelines and creates friction in the customer experience compared to more automated competitors.

03 Regulatory Compliance Concerns

The Equal Credit Opportunity Act requires non-discriminatory, verifiable lending decisions with clear justification for adverse actions. Manual processes make compliance documentation more difficult and increase risk of inconsistent treatment.

04 Suboptimal Resource Allocation

Loan officers currently spend significant time on routine application reviews instead of complex cases requiring judgment. This represents an opportunity cost where skilled personnel could be better utilized for relationship building and handling exceptions.

METHODOLOGY & APPROACH

Data Assessment & Preparation

- Comprehensive data quality analysis
- Treated outlier values while preserving legitimate extreme values
- Applied domain-specific imputation strategies for missing values

Exploratory Analysis

- Conducted univariate, bivariate and mutlivariate analyses
- Identified feature relationships with default status, critical thresholds and interaction patterns

Model Development

- Prepared data for model training with appropriate encoding and scaling applied
- Trained multiple model types (Logistic Regression, Decision Tree, Random Forest & Gradient Boosting)
- Applied hyper-parameter tuning to improve model performance
- Benchmarked model performance using accuracy, precision, recall and F1 scores

DATA INSIGHTS

Credit History

Delinquent credit lines and derogatory reports show 427% and 401% higher values, respectively, in defaulting loans

Financial Thresholds

Default risk triples from 5.7% to 17.8% when exceeding a debt-to-income ratio of 39%

Occupation-Purpose Interaction

Self-employed borrowers seeking debt consolidation show 43.8% default rate

Credit Maturity

Each quintile of credit age reduces default risk by ~5%, dropping from 30% default rate for newest credit histories to just 10% for the most established accounts

SOLUTION DESIGN

MODEL SELECTION

Random Forest Classifier

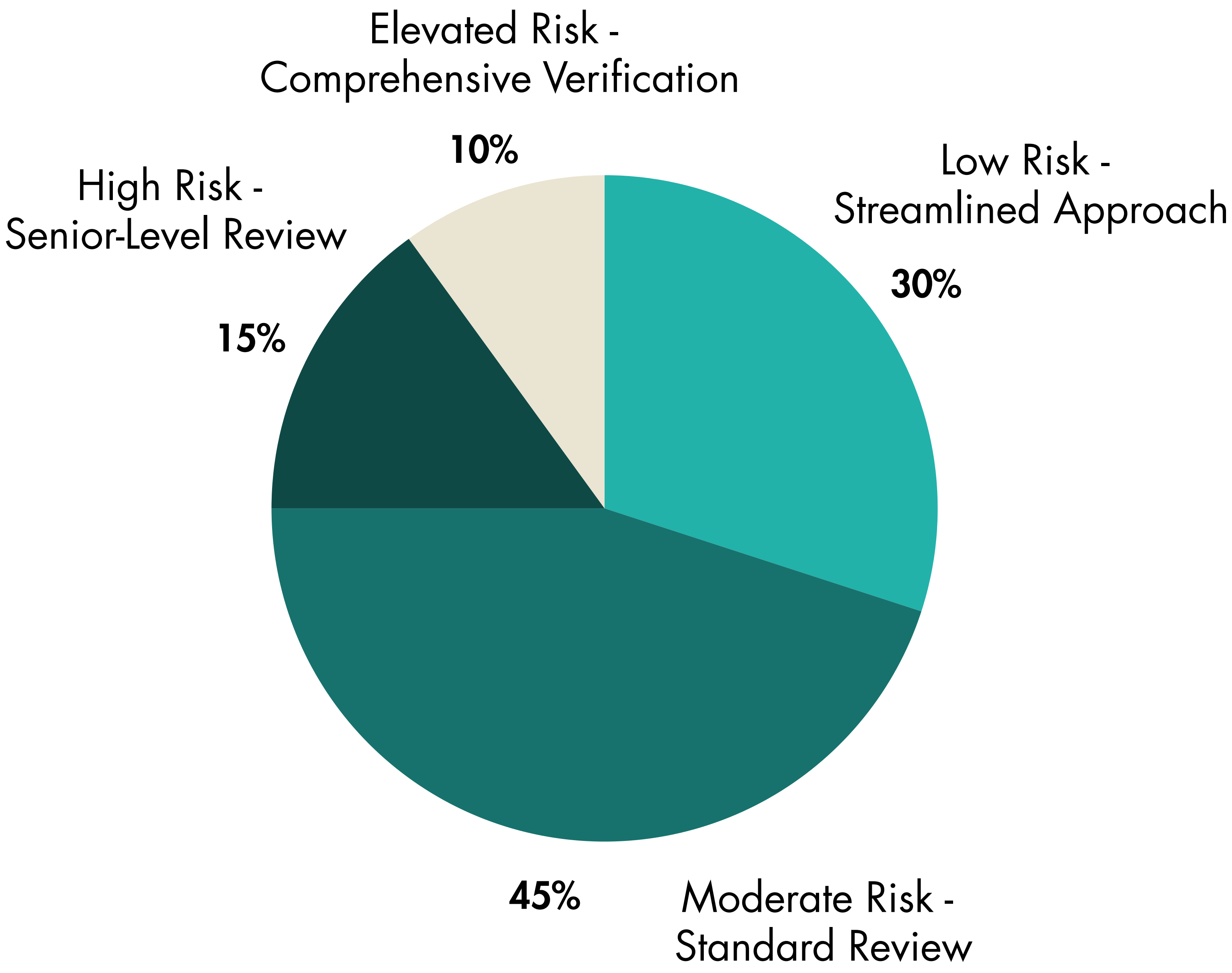
- 91% overall accuracy
- Primary decision engine for balanced precision-recall

Optimised Decision Tree

- Captures 86% of defaults
- Simple, interpretable structure for regulatory compliance
- Secondary screening for high-risk detection

SOLUTION DESIGN

IMPLEMENTATION FRAMEWORK



01 Risk-Based Approval Tiers

Categorization of applications into four risk levels, each with appropriate handling

02 Specialized Handling for High-Risk Segments

Development of targeted approaches for specific high-risk segments identified in through analysis

03 Four-Phase Implementation Plan

To ensure seamless integration and optimal performance, implementation of this model will be done in phases

EXPECTED BENEFITS

Financial Impact

- Decreased processing cost through automation
- Improved revenue through faster, more accurate decisions

Operational Improvements

- Consistent, data-driven approval process
- Better resource allocation for loan officers
- Enhanced regulatory compliance through documentation

20%	Assuming a conservative 20% reduction in loan defaults
$4.4M	Assuming a constant loan amount equal to the average, saves the bank $4.4M

IMPLEMENTATION ROADMAP

Phase 1 - Parallel Testing

Running models alongside existing processes without impacting decisions for comparison and discrepancy identification

Phase 2 - Pilot Deployment

Implementing the system for a controlled subset of applicants in order to monitor performance and collect feedback

Phase 3 - Full Implementation

Expand model application to all loans, alongside dashboards for long term monitoring

Phase 4 - Ongoing Refinement

Regular model calibration and development of enhanced analytics based on implementation learnings

Great Learning Capstone Project

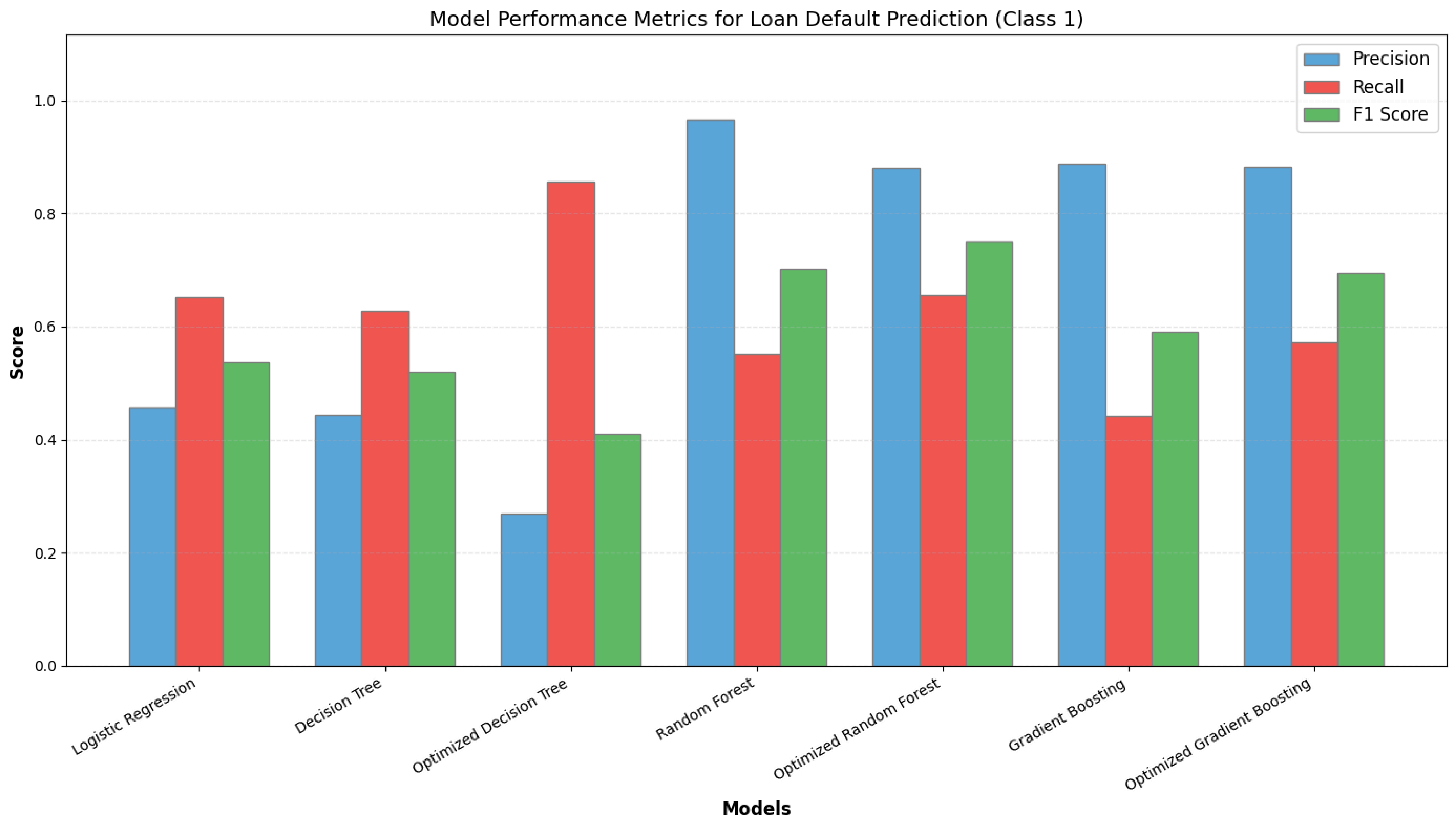
THANK YOU

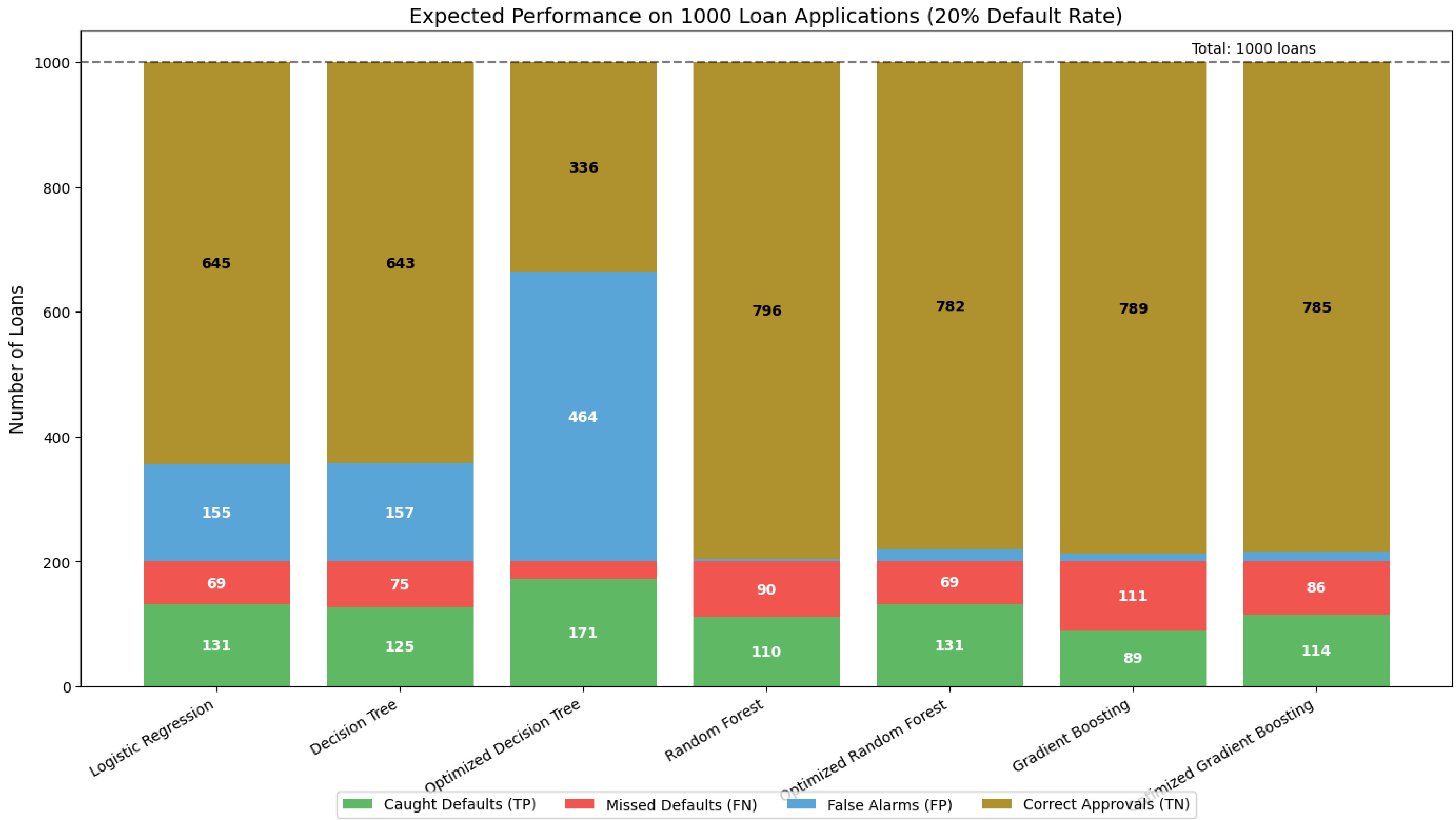
FOR YOUR ATTENTION

Presented By
Natasha Soldin

April 2025

APPENDIX


Model Performance Metrics for Loan Default Prediction (Class 1)
Precision
Recall
F1 Score
Score
1.0
0.8
0.6
0.4
0.2
0.0
Logistic Regression
Decision Tree
Optimized Decision Tree
Random Forest
Optimized Random Forest
Gradient Boosting
Optimized Gradient Boosting
Models



Expected Performance on 1000 Loan Applications (20% Default Rate)
Total: 1000 loans
Number of Loans
0
200
400
600
800
1000
645
155
69
131
643
157
75
125
336
464
171
796
90
110
782
69
131
789
111
89
785
86
114
Logistic Regression
Decision Tree
Optimized Decision Tree
Random Forest
Optimized Random Forest
Gradient Boosting
Optimized Gradient Boosting
Caught Defaults (TP)
Missed Defaults (FN)
False Alarms (FP)
Correct Approvals (TN)

Logistic Regression:

Missed Defaults Cost: $3,473,389 (69 loans at $50,000 each)

False Alarm Cost: $774,284 (155 loans at $5,000 each)

Total Cost: $4,247,673

Decision Tree:

Missed Defaults Cost: $3,725,490 (75 loans at $50,000 each)

False Alarm Cost: $782,669 (157 loans at $5,000 each)

Total Cost: $4,508,160

Optimized Decision Tree:

Missed Defaults Cost: $1,428,571 (29 loans at $50,000 each)

False Alarm Cost: $2,320,056 (464 loans at $5,000 each)

Total Cost: $3,748,627

Random Forest:

Missed Defaults Cost: $4,481,793 (90 loans at $50,000 each)

False Alarm Cost: $19,567 (4 loans at $5,000 each)

Total Cost: $4,501,359

Optimized Random Forest:

Missed Defaults Cost: $3,445,378 (69 loans at $50,000 each)

False Alarm Cost: $89,448 (18 loans at $5,000 each)

Total Cost: $3,534,826

Gradient Boosting:

Missed Defaults Cost: $5,574,230 (111 loans at $50,000 each)

False Alarm Cost: $55,905 (11 loans at $5,000 each)

Total Cost: $5,630,135

Optimized Gradient Boosting:

Missed Defaults Cost: $4,285,714 (86 loans at $50,000 each)

False Alarm Cost: $75,472 (15 loans at $5,000 each)

Total Cost: $4,361,186

Lowest Cost Model: Optimized Random Forest

Total Cost: $3,534,826

This model provides the best financial outcome based on the assumed cost structure.