Capstone Project

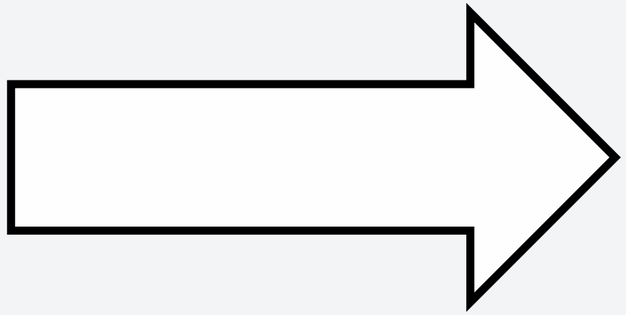
LOAN DEFAULT PREDICTION

Sprint 2

NOVEMBER 2023

Presented by:
David Clarke

AGENDA



1. Re-intro to Problem
2. EDA Insights
3. Data Preprocessing
4. Baseline Models
5. Next Steps

PROJECT OVERVIEW

Subject Area:

Predicting Loan Defaults based on borrower characteristics

Opportunity Identified:

Approx. **2.5%** delinquency rate of approx. **$17 trillion** consumer debt

⇨ **~$400 billion loss in financial markets**

Proposed Solution:

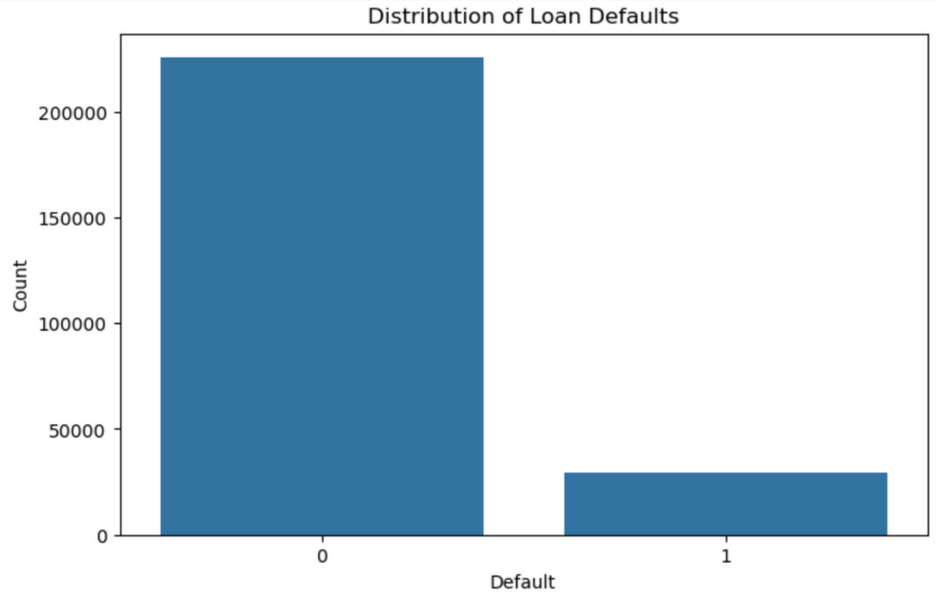
Leverage sophisticated ML algorithms to improve accuracy

⇨ **1% improvement = ~$4 billion savings**

EDA FINDINGS

Major Concerns:

- Significant DATA IMBALANCE in the target variable

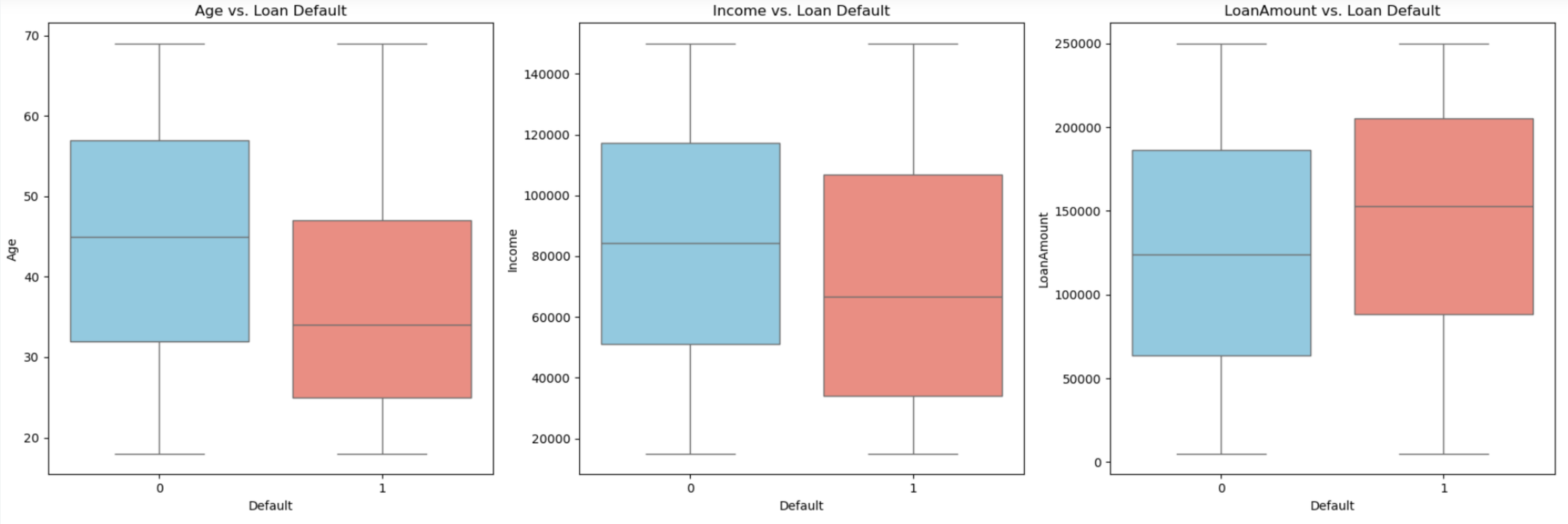


EDA FINDINGS

Action Plan:

- Resampling Techniques
- Algorithmic Approaches
- Evaluation Metrics (precision, recall)
- Ensemble Methods (bagging - RF, boosting - GBM)
- Hyperparameter tuning

EDA FINDINGS



DATA PREPROCESSING

Techniques considered:

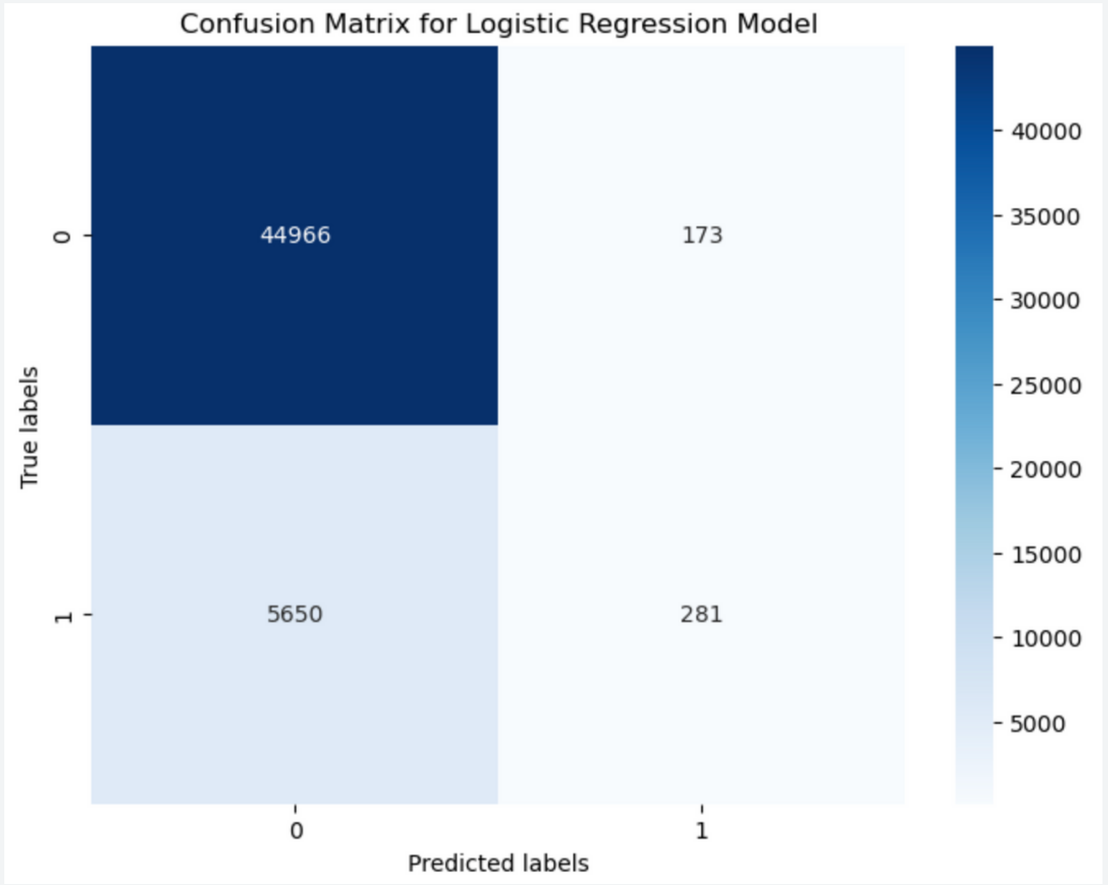
- **Binning Continuous Variables**
  - Age -> "Young", "Middle-aged", "Senior".
  - Income levels -> "Low", "Medium", "High".
  - CreditScore -> "Poor", "Good", "Excellent".
- **One-hot Encoding**
- **Creating New Features**
  - Analyzing interactions between features such as Income and Loan Amount
- **Feature Scaling**

BASELINE MODELING

Logistic Regression:

- Train acc: 88.5%
- Test acc: 88.5%

	precision	recall	f1-score	support
0	0.89	1.00	0.94	45139
1	0.62	0.05	0.09	5931
accuracy			0.89	51070
macro avg	0.75	0.52	0.51	51070
weighted avg	0.86	0.89	0.84	51070

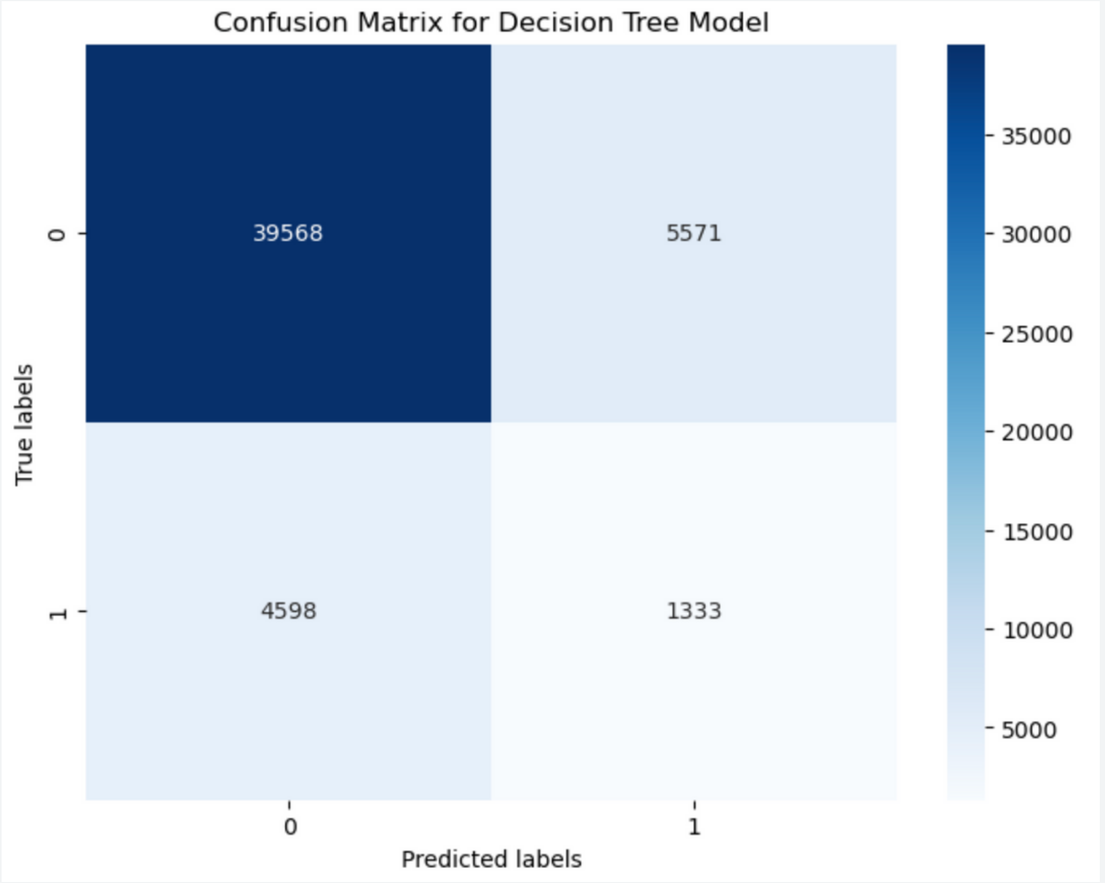


BASELINE MODELING

Decision Tree:

- Train acc: 100%
- Test acc: 80%

	precision	recall	f1-score	support
0	0.90	0.88	0.89	45139
1	0.19	0.22	0.21	5931
accuracy			0.80	51070
macro avg	0.54	0.55	0.55	51070
weighted avg	0.81	0.80	0.81	51070



NEXT STEPS:

1) Advanced Modeling techniques:

- Random Forests, Gradient Boosting Machines, SVM's
- Model tuning: Grid search and Hyperparameter optimization
- Reassessing feature selection

2) Evaluation Strategy Enhancement

- K-fold cross validation
- Alternative metrics

3) Model Deployment

- AWS/ local server

Thank you