

Loan Default Prediction

Practical Data Science - Capstone Project

Evita Delikoura

MIT-PE ADSP NOVEMBER 23'C

A dark blue diagonal gradient bar that starts from the bottom left corner and extends towards the top right corner, covering the lower half of the slide.

Agenda

1. Context & Problem Statement
2. Solution Design
3. Key Insights
4. Recommendations & Risks

Context & Problem Statement

Banks Make \$\$\$ → Repaid Home Loans

Problem Today: To optimize business and reduce financial loss, banks must be able to accurately predict defaulters



PROPOSED SOLUTION: Build predictive classification model to minimize acceptance error of defaulters

IDENTIFY:

- Key drivers of default
- Important features to consider when approve loan
- What type of applicant is likely to default

Data Overview

Home Equity Dataset (HMEQ) -
5,960 recent home equity loan data,
13 variables

- Current Approval Process =
\$20 million loss
- 80/20 data split

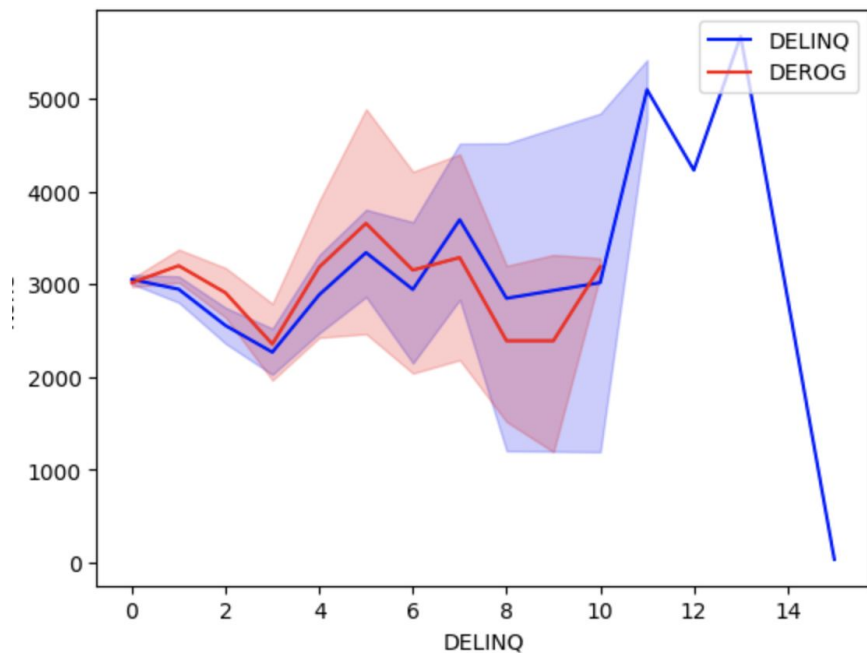
Numerical & Categorical variables

All variables have missing values,
except LOAN and target BAD

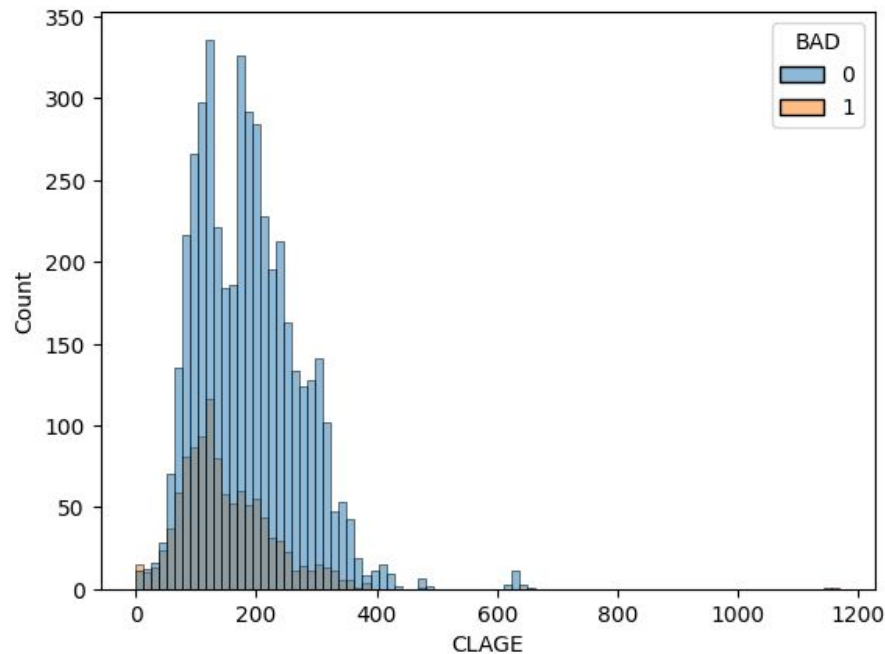
#1 Reason: Debt Consolidation,
\$19K loan average

Variable	count	mean	std	min	25%	50%	75%	max
BAD	5960	0.2	0.4	0.0	0.0	0.0	0.0	1.0
LOAN	5960	\$18,608	\$11,207	\$1,100	\$11,100	\$16,300	\$23,300	\$89,900
MORTDUE	5442	\$73,761	\$44,458	\$2,063	\$46,276	\$65,019	\$91,488	\$399,550
VALUE	5848	\$101,776	\$57,386	\$8,000	\$66,076	\$89,236	\$119,824	\$855,909
DEROG	5252	0.3	0.8	0.0	0.0	0.0	0.0	10.0
DELINQ	5380	0.4	1.1	0.0	0.0	0.0	0.0	15.0
CLAGE	5652	179.8	85.8	0.0	115.1	173.5	231.6	1168.2
CLNO	5738	21.3	10.1	0.0	15.0	20.0	26.0	71.0
DEBTINC	4693	33.8	8.6	0.5	29.1	34.8	39.0	203.3

Insights from EDA



DELINQ & DEROG = 28-34% correlation to BAD




The older the credit lines, the less likely of default

Data Preparation

- Filled missing values with median and mode
- Create Missing value flag
- Replaced outliers with maximum
- One hot encoded categorical variables
- Split data sets (imbalanced) to train and test
 - Class 0: 0.20
 - Class 1: 0.80

Model Building Comparison

Model	Class	Precision	Recall	F1-score	Accuracy
Logistic Regression	0	0.81	1.00	0.89	0.81
	1	0.78	0.04	0.07	
Decision Tree	0	0.91	0.92	0.91	0.86
	1	0.66	0.63	0.65	
Tuned Decision Tree	0	0.93	0.90	0.91	0.87
	1	0.65	0.71	0.68	
Random Forest	0	0.92	0.96	0.94	0.90
	1	0.80	0.64	0.71	
Tuned Random Forest 	0	0.93	0.93	0.93	0.89
	1	0.73	0.73	0.73	

PARAMETERS for BEST MODEL:

1. Maximizes Recall
2. Hypertuned: Class weights, GridSearch
3. Accounts for all Features

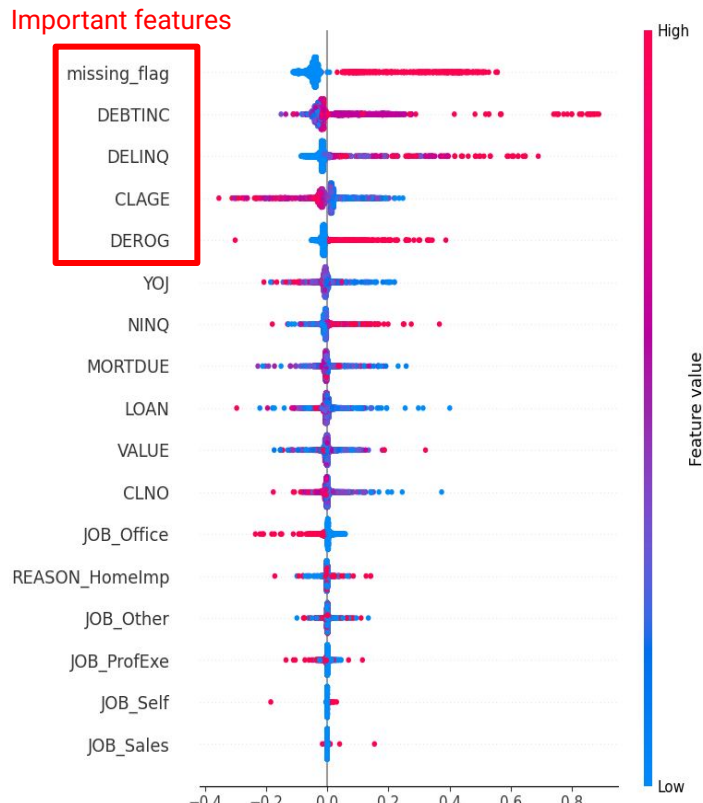
Best Model: Tuned Random Forest Classifier

Model Performance

- Maximizes Recall: Class 1 73%, Class 0 93%
- High Accuracy: 89%
- Hypertuned using Class Weights + GridSearch
- Captures all features
- Interpretable

Key Features

- Missing flag
- DEBTINC
- DELINQ
- CLAGE



Recommendations & Risks

In order to deploy model into production:

1. Make all input fields required
2. Gather more customer information
3. Monitor & maintain performance of model

Risks:

1. Outcome of default
2. Trade off between financial loss vs opportunity cost