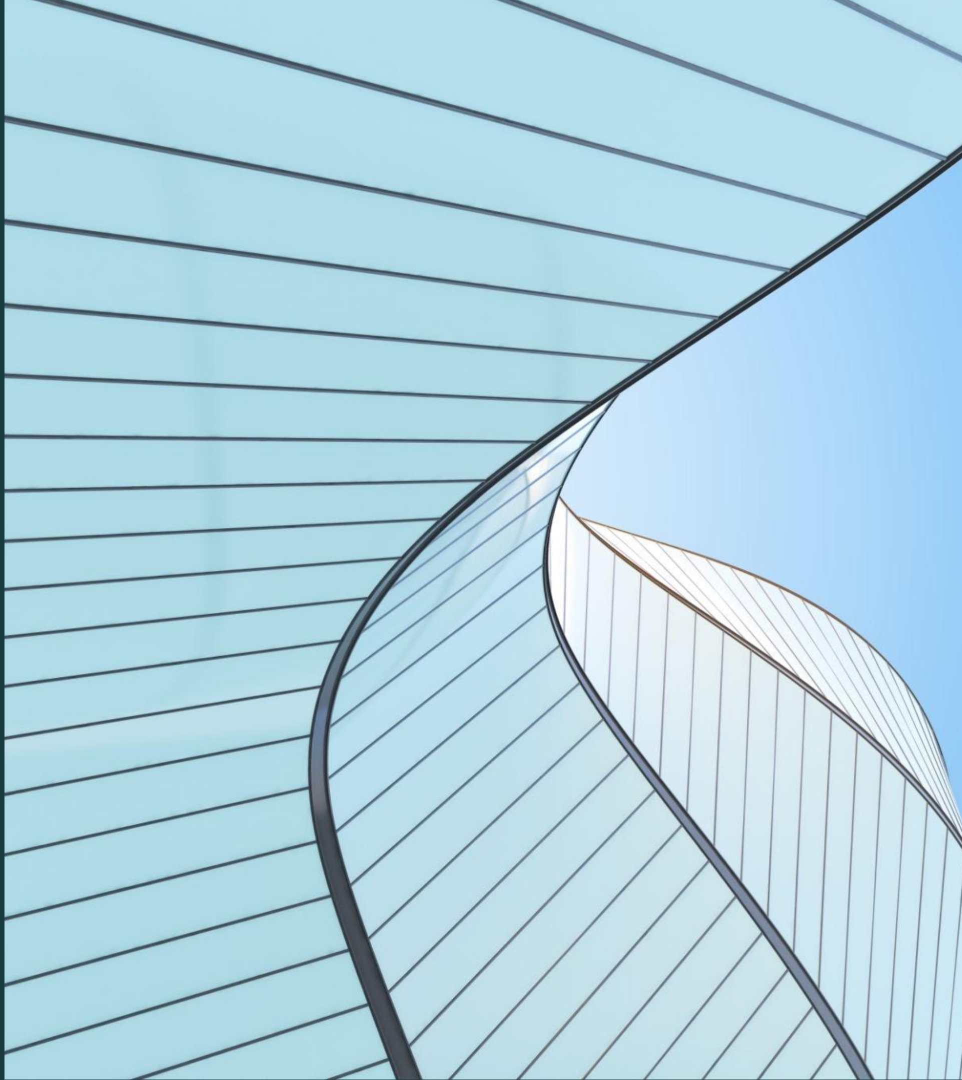
Loan Default Prediction

Home Equity Loan Default Prediction following the guidelines in the Equal Credit Opportunity Act



Key Terms

Home Equity Loan → A fixed-amount loan taken out against the equity in a client's home.

Equity → Defined as the amount a client can borrow against their home.

Debt to Income Ratio → Defined as the ratio of a client's monthly debt to their gross income.

Delinquent Credit Line → A line of credit where the client has not made the minimum payment required for more than 30 to 60 days.

Problem Overview

Interest on Home Equity Loans is a ***major*** portion of a bank's profit.

If a client default on their loan, the bank ***cannot*** turn a profit on that loan.

Determining whether or not a client is going to default on their loan is **crucial** to business operation.

Problem Solution

The bank's Consumer Credit Department would like to establish a **simplified decision making model** that predicts client's who will default on their home equity loan (HEQ).

Requirements

Follow the guidelines in the Equal Credit Opportunity Act (HMEQ).

Use the data provided from the *existing* loan underwriting process

Solution Approach

Part One

Understand the structure and components of the data.

Part Two

Explore the data using to gain key insights

Part 3

Create and tune 3 different classification models.

Part 4

Select the model that best fits the data and aligns with the desires of the Consumer Credit Department

Our goal was to analyze the data given to us by the Consumer Credit Department, then use our analysis as a guide *to build a model that will simplify the decision making process for lending Home Equity Loans.*

Data Breakdown

The dataset provided contains information recently dispersed Home Equity Loan records.

01

There are 5960 records.
Each record is described by 13 features.

02

Our target is inscribed to the feature ***Defaulted***.

03

Defaulted contains two classes:

1 = Client defaulted on loan

0 = Client repaid loan

04

Key Takeaways

Debt to Income Ratio

Clients whose debt to income ratio is *larger* on average were more likely to **default**.

01

Delinquent Credit Lines

Clients who **defaulted** on average *had more* delinquent credit lines than those who didn't.

They also had more derogatory reports.

02

Loan Amount

On average, clients who **didn't default** were *given higher valued* Home Equity Loans than those who did.

These clients also had higher property values and mortgages.due.

03

Professions

A client's who listed their profession as *Sales* were more likely to **default** than all other professions.

However, *Other* is the largest listed profession and the second most likely to **default.**

04

Proposed Solution Models

Logistic Regression
Gives us the predicted probability that a client will either default or repay their loan.

Classification Decision Tree
The Classification Decision Tree will classify a client as one of two people, those will default or those who will repay their loan.

Random Forest
Many Decision Trees are made via Bootstrapping the data and Aggregating their results into a final classification of the client using the same target as the Classification Decision Tree.

Decision Tree

Visibility

The graphical representation of a Decision Tree provides a flow chart to follow the decision making process.

Key Feature Selection

A hierarchy of feature relevance is created that is easily interpretable and naturally de-emphasizes irrelevant features.

Recall

The model that is able to maximize Recall will also be the one who will minimize predicting the false negative of a client who *will repay* their loan but in reality *will default* on the loan.

Executive Summary

Problem

The Consumer Credit Department would like to automate and simplify the Home Equity Loan decision making process via utilizing Data Science and Machine Learning.

Key Features

1. Debt to Income Ratio
2. Number of Delinquent Lines
3. Age of the Oldest Credit Line
4. Loan Amount

Proposed Model

Classification Decision Tree

Proposed Business Solution

01

Prioritize Debt to Income Ratio

a client with a *high debt_to_income ratio* will be more likely to default than one who has a lower ratio.

02

Strongly Consider Delinquent Lines

the *number of delinquent credit lines* directly reflects on their ability to take on another credit line such a HEQ loan.

03

Focus on Credit Age

the *age of a client oldest credit line* reflects positively on their relative trustworthiness with a new line.

04

Minimally to Moderately Consider the Following

1. Years at Current Job
2. Reason for Loan Request
3. Profession
4. Loan Request Amount

Proposed Model Improvements

01

Balance Out the Clients

Adding *additional clients who defaulted* into the data to provide a more balanced view of the causes of default.

02

Prioritize Filling Missing Values

Client records missing the key features of *debt to income ratio*, number of *delinquent lines*, and the *age of the oldest credit line*.

03

Add Equity

Consider the potential benefits of combining the values of *mortgage due* with *property value* via calculating *equity*.

04

Consider Combining Correlated Features

Different *combinations* of moderate to highly correlated features improve the model's ability to make decisions.

Risks and Challenges

- No model is perfect and errors can happen.
  - Decision Comprehension
- A Decision Tree is *Moderately Robust*.
  - Outliers & Diversity
  - Small Changes
- Focusing on a few *Key Features* could lead to unintentional bias.

Appendix

The original dataset had ambiguous labels for the features so they were updated to improve understandability.

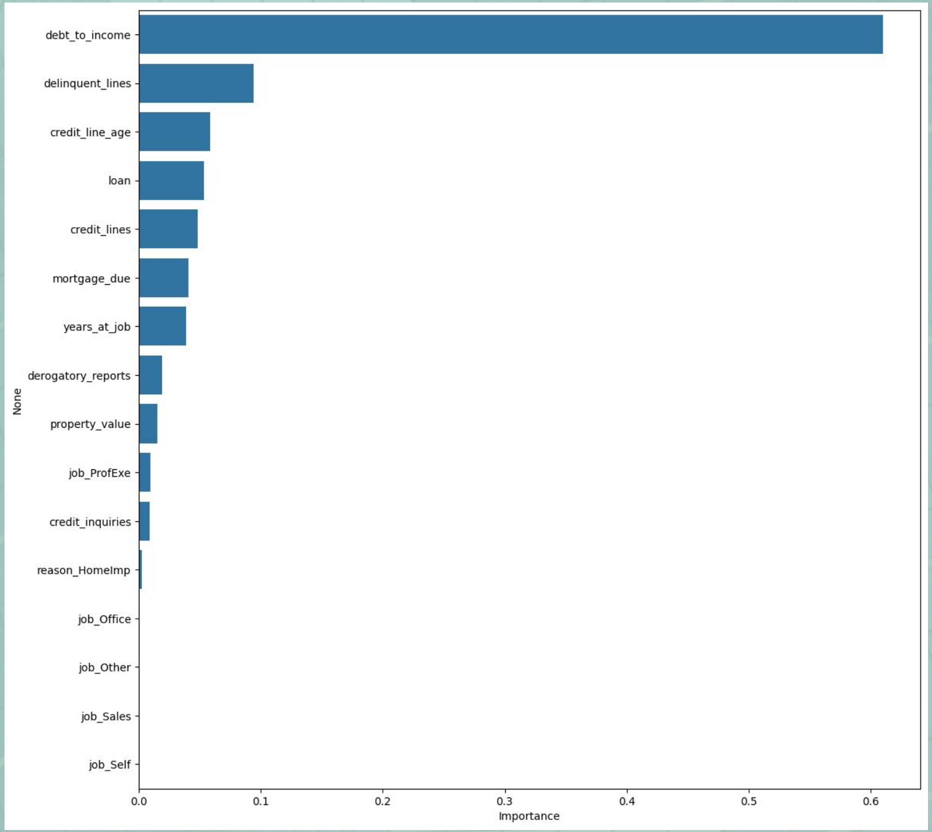
Model Comparison

	RECALL	PRECISION	ACCURACY	SPECIFICATIONS
Tuned Logistic Regression	• Train = 0.45 • Test = 0.33	• Train = 0.68 • Test = 0.50	• Train = 0.81 • Test = 0.80	• Threshold = 0.35
Tuned Decision Tree	• Train = 0.87 • Test = 0.83	• Train = 0.59 • Test = 0.56	• Train = 0.85 • Test = 0.83	• class weight = 'balanced' • impurity measure = 'entropy' • max depth = 9 • min samples per leaf = 25
Tuned Random Forest	• Train = 0.79 • Test = 1.00	• Train = 0.63 • Test = 0.60	• Train = 0.86 • Test = 0.87	• number of estimators = 120 • class weight = 'balanced' • Impurity measure = 'entropy' • max depth = 5 • min samples per leaf = 20

Logistic Regression Feature Importance

	odds
delinquent_lines	1.887727
credit_inquiries	1.447648
derogatory_reports	1.381896
profession_Self	1.020849
reason_HomeImp	1.015627
profession_Sales	1.012631
debt_to_income	1.006292
property_value	1.000002
mortgage_due	0.999995
loan	0.999973
credit_line_age	0.992729
profession_Other	0.989146
profession_ProfExe	0.985828
years_at_job	0.980178
credit_lines	0.978594
profession_Office	0.938289

Decision Tree Feature Importance



Random Forest Feature Importance