# Predicting Loan Defaults

Christopher Campbell

# Key Points

- Problem:  Loan Defaults
- Solution:  Machine Learning techniques applied to the problem
- Best  model?
- Insights from the model: Key drivers of default
- Recommendations

# The Problem

- Bad loans (loans that default)  are a major source of risk for lenders since they eat away at profits and therefore, it is important for Banks to ensure that they don't approve loans that are likely to default.
- The traditional loan approval process is labor intensive and prone to error,  requiring loan officers and underwriters to manually review  documents from various sources.
- With this is mind, we are to build a predictive model that will simplify the decision making process for rejecting loans.
- More specifically , we will build a **classification model** to predict clients who are likely to default on their loan.

# Solution: Classification Models

- Logistic Regression
- Decision Tree
- Random Forest

These are standard modeling techniques  that are statistically sound and utilized in various regulatory environments

# The Data

The Home Equity dataset (HMEQ) contains baseline and loan performance information for 5,960 recent home equity loans. The target (BAD) is a binary variable that indicates whether an applicant has ultimately defaulted or has been severely delinquent. This adverse outcome occurred in 1,189 cases (20 percent ). 12 input variables were registered for each applicant.

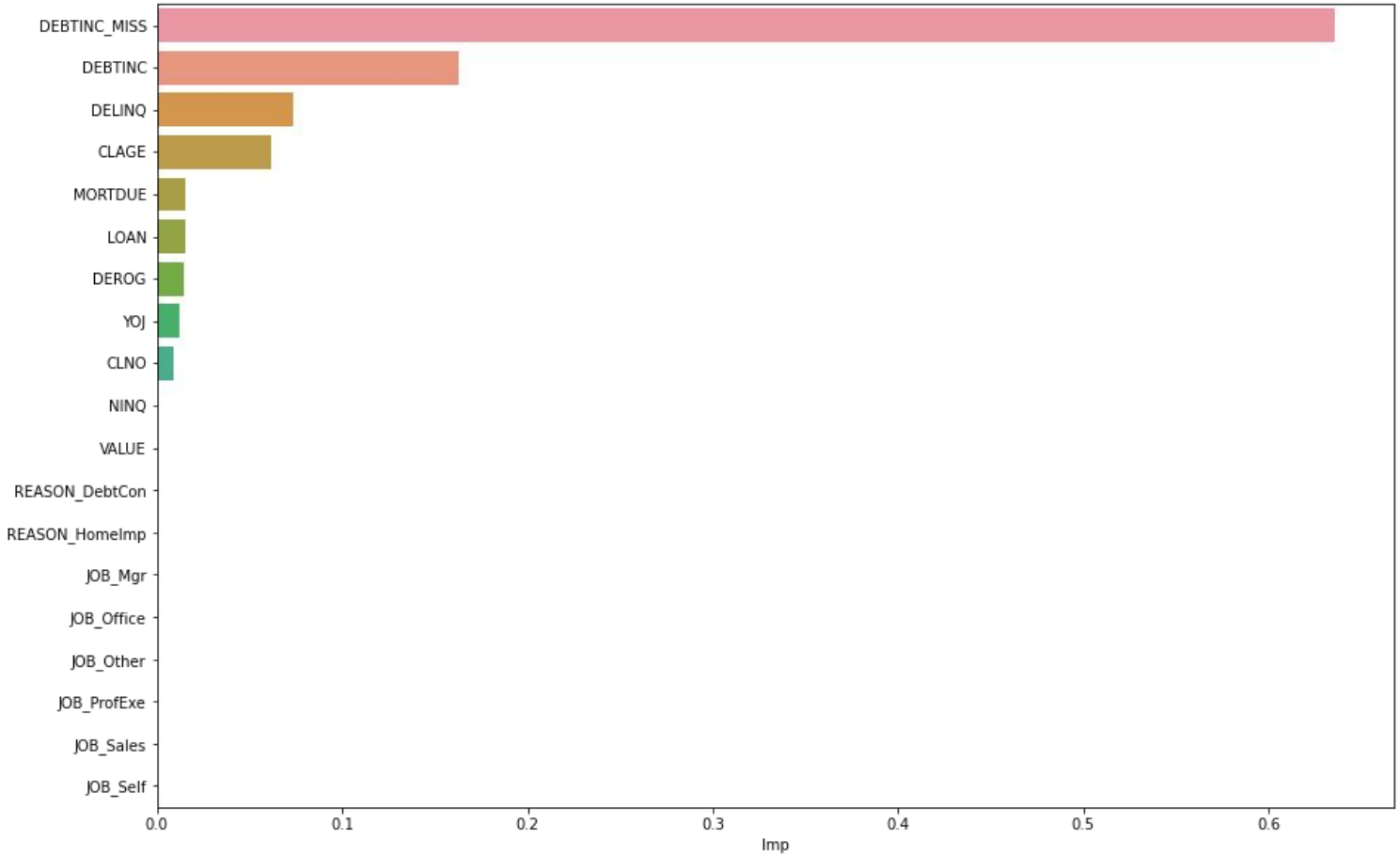
- BAD: 1 = Client defaulted on loan, 0 = loan repaid
- LOAN: Amount of loan approved.
- MORTDUE: Amount due on the existing mortgage.
- VALUE: Current value of the property.
- REASON: Reason for the loan request. (HomeImp = home improvement, DebtCon= debt consolidation which means taking out a new loan to pay off other liabilities and consumer debts)
- JOB: The type of job that loan applicant has such as manager, self, etc.
- YOJ: Years at present job.
- DEROG: Number of major derogatory reports (which indicates a serious delinquency or late payments).
- DELINQ: Number of delinquent credit lines (a line of credit becomes delinquent when a borrower does not make the minimum required payments 30 to 60 days past the day on which the payments were due).
- CLAGE: Age of the oldest credit line in months.
- NINQ: Number of recent credit inquiries.
- CLNO: Number of existing credit lines.
- DEBTINC: Debt-to-income ratio (all your monthly debt payments divided by your gross monthly income. This number is one way lenders measure your ability to manage the monthly payments to repay the money you plan to borrow.

# Solution: Decision Tree

We determined that the best model to be used is Tuned Decision Tree.  Key strengths of this model are:

1. Good (Not Best) Performance:  The most important metric for this problem,  correctly identifying which loans will default:  Model - 81% of customers that actually default will be correctly labeled as defaulters.
2. Stable:  Performance is similar across key metrics and  across train and test data sets.
3. Interpretable and Understandable:  Worth noting that Random Forest performed slightly better than Decision tree.  However, Decision trees  have superior interpretability and explainability which could be of great value in light of potential legal and/or regulatory issues.

# Relative Importance of Customer Data Fields for Defaults



# Question:

If missing DEBTINC is a leading indicator of loan defaults, then why / what causes customers to have missing values for this data field??

# Recommendations

- Do NOT underwrite loans to customers that have missing DEBTINC data.  These almost always default.
- Do NOT underwrite loans to customers that have DEBTINC data but have DEBTINC ratios > 43.57.  These always default.
- Do NOT underwrite loans to customers that have DEBTINC data but have DEBTINC ratios <= 43.57 and DELINQ > 1.50.  These always default.

# Benefits of Implementation

- Implicit variable selection - top most variables in the tree are the most important
- Minimal data prep - data does not need to be normalized and decision trees less sensitive to outliers / missing data.
- Decision Tree output is graphical and easy to explain
- Screening could be done with minimal additional costs - reject customers with missing DEBTINC