



Loan Default Prediction

Capstone Project for MIT ADSP Jan 2022

Report by: Shweta S. Chandole

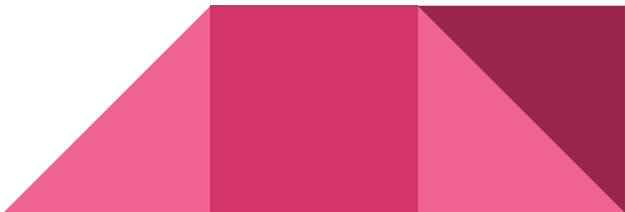
Outline

- Overview
- Data
- Solution
- Models
- Metrics
- Final Proposal
- Recommendations
- Q&A

Overview

- **Context**
- **Objectives**

Context

- Home equity loans (HMEQ) are consumer loans based on the guarantee of equity owned by the borrowers in their existing home value
 - At present Bank has manual loan approval process
 - 20% of approved loans have status BAD=1 i.e. Defaulted
 - With every defaulted loan, bank loses more money (loss of principal amount)
 - It is important to curtail this loss by stricter approval
 - Every defaulter predicted can save thousands in investment.
 - As per historic data, a total of \$110M loan amount was approved. Out of the approved loans, 20% were defaulted. The approx. loss of principal amount was \$20M, which is about 18% loss of investment.
- 

Objectives

- To build a predictive classification model to identify defaulters and non-defaulters
- To have better accuracy than manual process ($>80\%$)
- To avoid bias towards majority class
- To identify important independent features that help in classification of loan applicants as defaulters or non-defaulters based on the predictive model

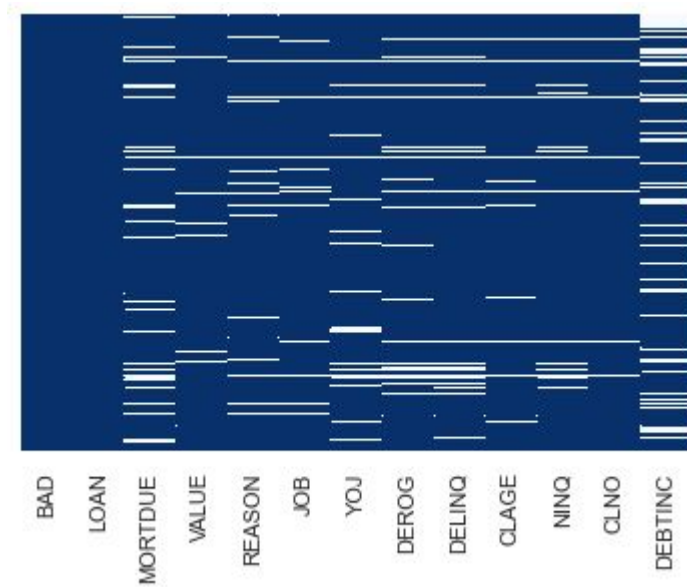


Data

- **Source Data**
- **Data Variables**
- **Challenges in the data**
- **Data Pre-processing**

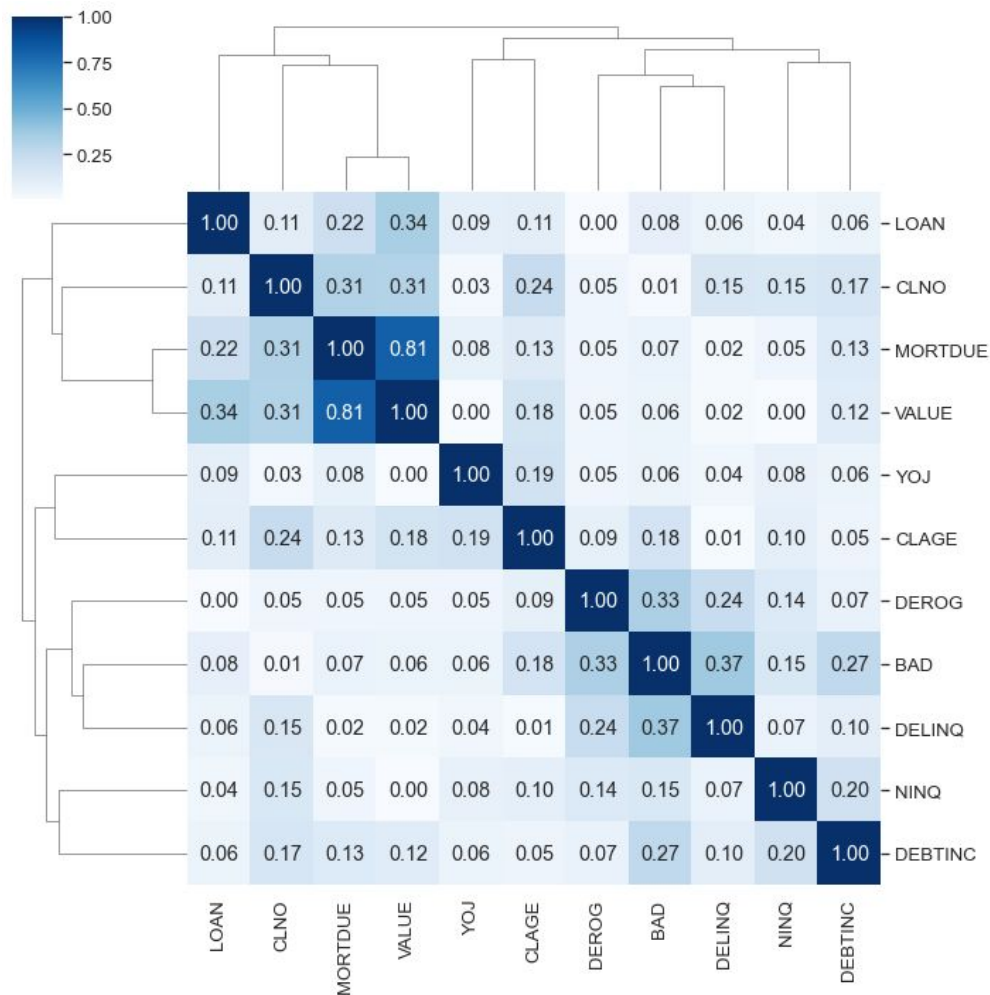
Source Data

- HMEQ dataset has 5960 unique records
- Target variable 'BAD' has 2 classes - 0 and 1
- LOAN values range from \$1100 to \$89,900
- High percentage data is missing in more than 60% features
- Image shows missing values in the columns of the dataset (Blue color indicates data is present, white color indicates data missing)



Data Variables

- Image shows correlation of various numerical variables in the data
- This helps in analysis of how each of the variable may be driving predictive power on the target variable



Challenges in Data

- Imbalanced target classes: 80% PAID (class 0), 20% DEFAULTED (class 1)
- Presence of outliers in numerical variables
- High percentage of missing values
- Skewed data distribution
- Presence of outliers
- Multicollinearity



Pre-processing of Data

- Data Cleansing
- Treating outliers before regression model
- Imputation of missing values with appropriate values per feature



Solution

- **Design Approach**
- **Performance Tuning**

Design approach

Data Preparation

- Train Test split using stratification
- Check for multicollinearity
- Oversampling of minority class using SMOTE technique for Logistic regression model


Classification models

- Logistic Regression
- Decision Tree
- Random Forest

Model Analysis

- Classification report
- Confusion matrix
- Feature importances

Performance tuning techniques

- Stratify train-test dataset splitting over the y-variable (binary stratification)
 - Optimal threshold of Precision-Recall curve
 - Resampling of training data using SMOTE
 - Class-weights in decision tree and random forest models
 - GridsearchCV using 5-fold Cross-validation for hyper-parameter tuning of Decision Tree model and Random Forest model
 - Feature Engineering to include new logical feature 'STATUS' and to transform discrete categorical features 'REASON', 'JOB', and continuous categorical features "DEROG" and 'DELINQ'
- 

Models

- **Logistic Regression**
- **Decision Tree**
- **Random Forest**

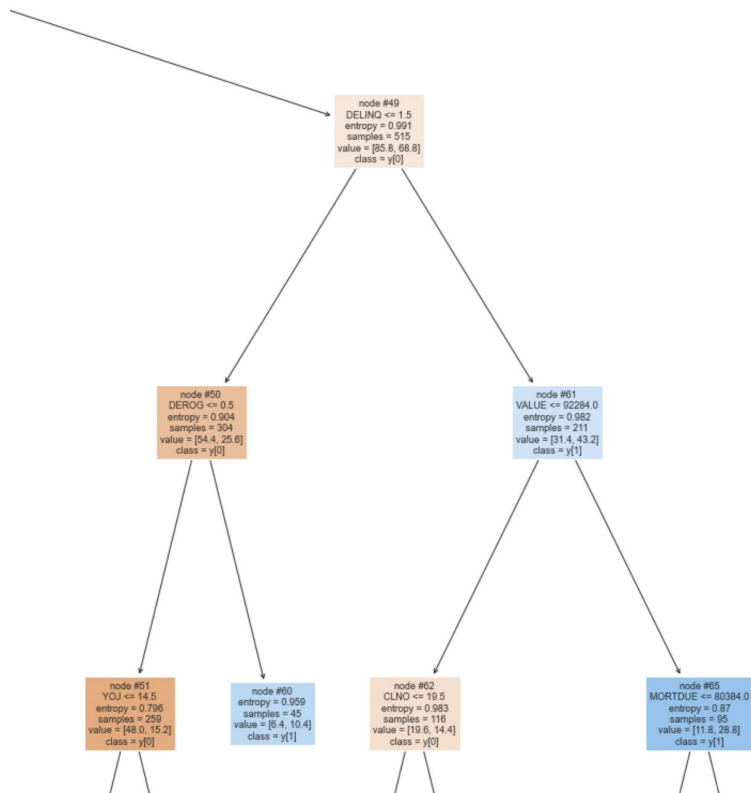
Logistic Regression Model

- Popular binary classification technique
- Indicates odds of an event happening using *logit*
- Sensitive to outliers - so data treated for outliers
- Model versions
 - ◆ Baseline model: poor recall on class 1
 - ◆ Precision Recall Curve to find optimal threshold for class 1: slight improvement
 - ◆ Resampling technique SMOTE used to balance minority class: slight improvement
- No significant results towards desired objectives



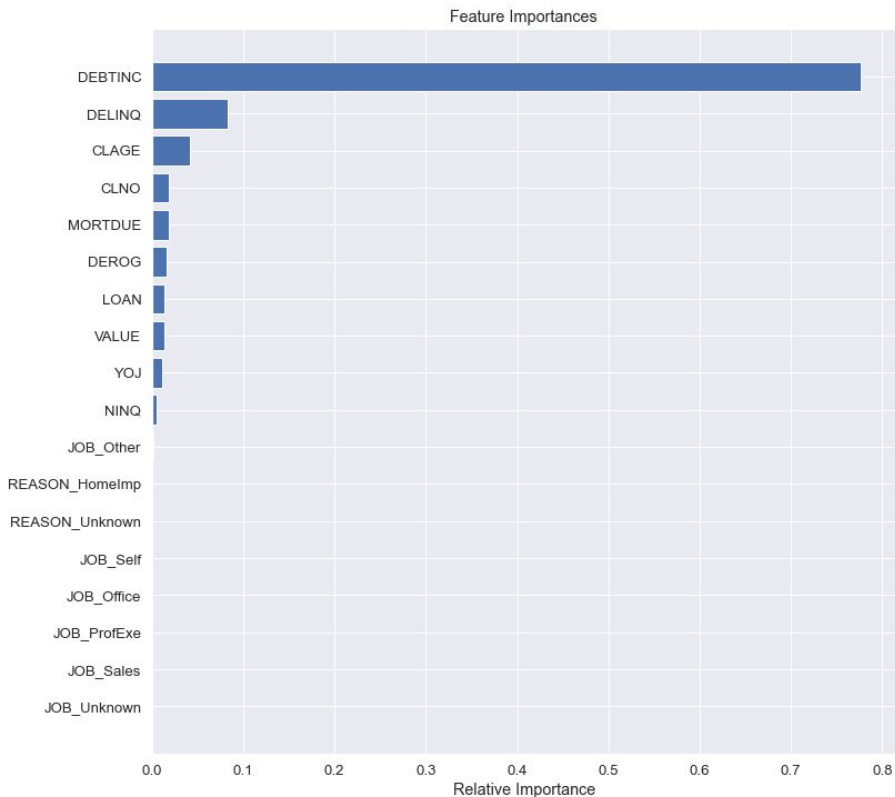
Decision Tree Model

- Supervised learning algorithm
- Visual display of decisions
- Robust to outliers
- Model versions
 - ◆ Baseline model: overfitted on training data
 - ◆ Tuned model: very good performance
- Important features DEBTINC, DELINQ, CLAGE, MORTDUE



Random Forest Model

- Supervised learning algorithm
- Ensemble learning method
- Cannot be visualized
- Robust to outliers
- Model versions
 - ◆ Baseline model: overfitted on training data
 - ◆ Tuned model: excellent performance
- Important features DEBTINC, DELINQ, CLAGE, CLNO



Metrics

- **Performance metrics**
- **Models comparison**

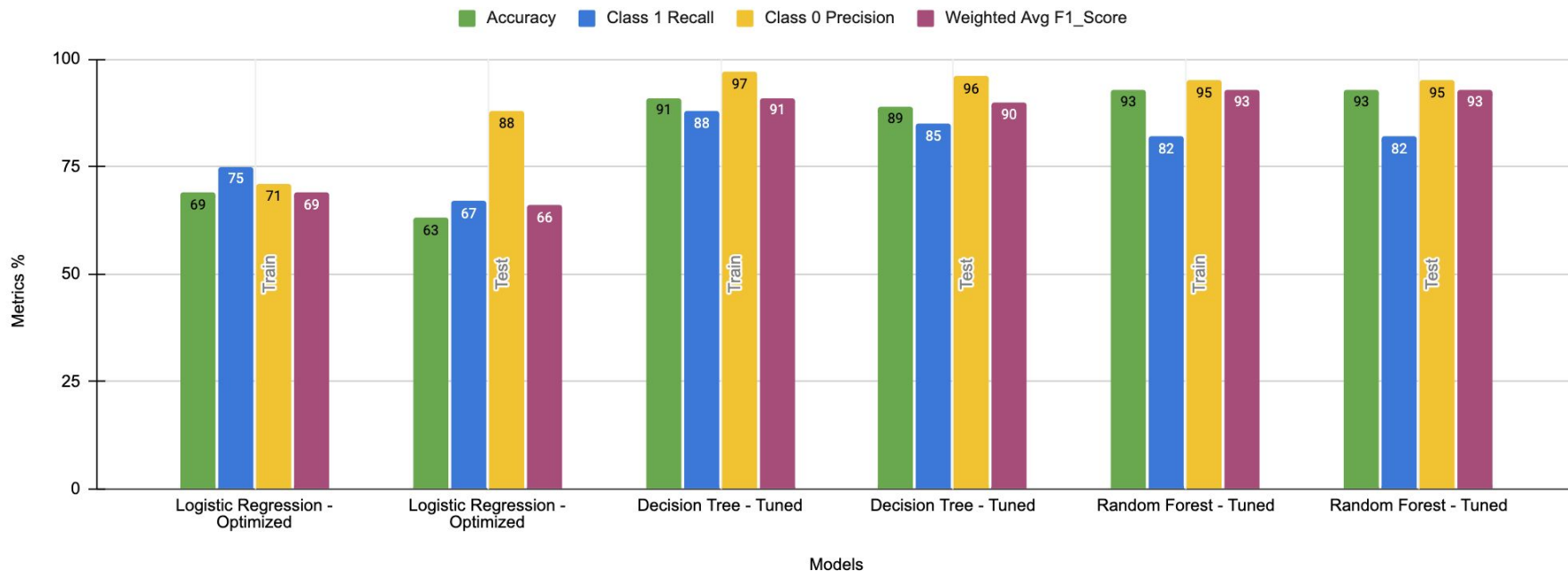
Performance Metrics

Evaluation Criteria	Metrics Score	Description
Maximize prediction of probable Defaulters	Recall Score for Class 1	Retrieval of most relevant records from data
Identifying Non-Defaulters with least error	Precision for Class 0	Relevant records out of retrieved ones
Balance rejection of Defaulters and acceptance of Non-Defaulters	Weighted Average F1 Score	Weighted sum of precision and recall for both classes
Make Accurate prediction of both classes	Accuracy	Accuracy of prediction for both classes

Models comparison

Loan Default Prediction Models

Comparison of Performance Metrics



Final Proposal

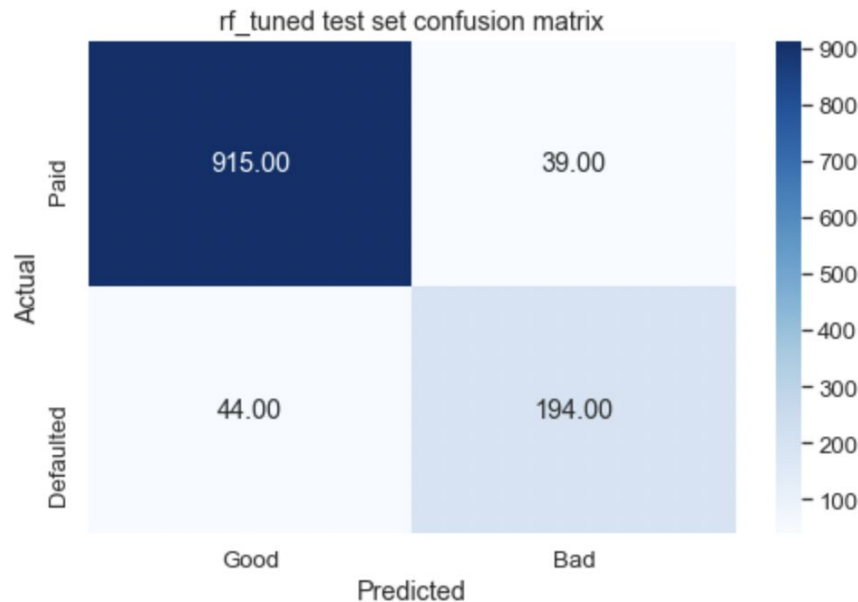
- **Proposed Model**
- **Cost Benefits**

Proposed Model

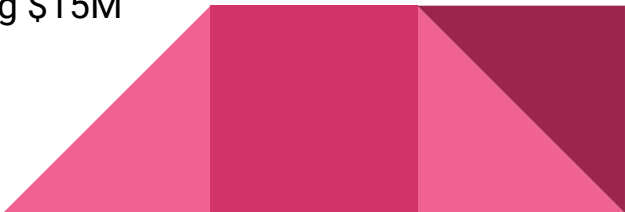
Based on the model analysis and comparison, we can conclude that the **Tuned Random Forest** model is the best model to meet the goals of

- 1) Model is fitted well on training data and shows same high-performance on test data
- 2) Maximize prediction of Defaulters with high class 1 recall rate of 82%
- 3) Balanced outcome of classification (highest true positive for both classes) F1-Score 93%
- 4) Highest accuracy of 93%
- 5) Key predictor variables are identified in prior discussion

rf_tuned test set metrics	set metrics			
	precision	recall	f1-score	support
0	0.95	0.96	0.96	954
1	0.83	0.82	0.82	238
accuracy			0.93	1192
macro avg	0.89	0.89	0.89	1192
weighted avg	0.93	0.93	0.93	1192



Benefits of the proposed solution

- With 20% defaulters in current data, approx. loss of principal amount was \$20M, which is about 18% loss of total approved loans amount \$110M
 - With the proposed solution, the best performing classification model can predict 82% defaulters with 93% accuracy
 - 95% Precision for identifying the non-defaulters
 - If we apply this predictive power to the historic data, we could identify $(0.93 \times 0.82 \times 20) = 15.25\%$ of defaulters with confidence, while only losing 5% of probable non-defaulters
 - With the probable defaulters brought down to 4.25%, the bank could reduce loss in principle amount by more than 75% amounting to less than \$5M approximately, saving \$15M
- 

Recommendations

Recommendations

- Estimated reduction in loss is 75%. Business could review the cost benefits of replacing existing approval process with the proposed solution on real time data
- Manual approval process could be replaced by a real-time automated process, with potential to process a higher number of applications with greater accuracy
- Solution design can be further enhanced to include features like suggest approval for partial loan amounts, or loan at interest rates in proportion to the credit risk of the loan applicant
- Better insights into the consumer's credit profile by adding features such credit score, income range, interest rates being offered for the loan, could help for further analysis.
- Having a larger dataset, with more complete and valuable information across features would work in best interest of the business to get a more accurate machine learning algorithm.





Q&A