Loan Default Prediction

A case of classic binary classification

Project Report

Shweta Chandole

January 2022

Table of Contents

Executive Summary 3

What could this mean to the business? 3

Problem and Solution Summary 4

Problem summary 4

Key objectives 4

Challenges in Data 4

Solution design approach 5

Model comparison 6

Final Proposed Solution 7

Recommendations 8

Benefits of proposed solution 8

Data considerations 8

Technical considerations 8

Business Considerations 8

Executive Summary

Consumer loans are one of the major contributors of revenue for retail banks.

Home equity loans (HMEQ) are consumer loans based on the guarantee of equity owned by the borrowers in their existing home value. Financial gain from processing such loans, in the form of fees and interests charged on the principal loan amount is a high-yielding business opportunity for the Bank.

However, it also involves a high risk if the borrower is unable to make timely repayments or defaults on the loan. When there is a large pool of applicants, losing a good loan will not be as much of a loss as approving a bad loan, as the bank can end up losing more money on principal amount of a defaulted loan as compared to earning interest on a good one.

The perceived gain of earning financial rewards on a paid loan can be presumably surpassed by the loss of principal on a defaulted loan. So, it is critical for the bank to have best judgement of possible defaulters with minimal error while approving these loans. This can be a cumbersome, time-consuming process, with possibility of human error or biases, since it is based on manual loan approval process used by the bank.

From consumer's perspective, it might also be preferable to have easy and fast access to the loan approval process, with a short turnaround time.

Thus, the Bank could benefit from having an automated process of predicting likelihood of an applicant being Defaulter or aNon-Defaulter with high accuracy. This can also mean ability to process increased number of applications and higher business opportunity for the Bank.

The key objectives here would be to predict probability of whether a loan applicant is most likely to default (class 1) or not likely to default (class 0), and also to identify the key features that have most predictive power on making this classification.

With these goals in mind, I have developed a Classification model using machine learning techniques like regression and tree-based classifications, to help with key objectives.

What could this mean to the business?

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.

With the proposed solution, the best performing classification model can predict 82% defaulters with 93% accuracy. Also, it has 95% precision on identifying the non-defaulters. If we apply this predictive power to the historic data, we could identify (0.93*0.82*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. The saving of $15M could pay for itself even if some business was lost due to over-cautious dismissal.

Business could benefit from investing in acquiring high quality data, with additional information on consumer's credit profile like credit score, income range, interest rates to help with developing better fitted models.

Problem and Solution Summary

Problem summary

We study the loan status and aspects of credit history for 5960 unique observations of HMEQ data in the given dataset. We need to develop a classification model that is robust to anomalies in the data, and has good fit over the training data and is able to generalize well on the test data.

Key objectives

- To build a predictive machine learning model that classifies the loan applicant as most likely to be a defaulter or a non-defaulter, with increased accuracy as compared to the manual process.
- To maximize true positives rate for class 1 (defaulters) in order to avoid loss by inaccurate approvals (Highest Class 1 Recall score is most important)
- To check for removal of errors and bias towards any one of the two classes of target variable (thus weighted averages F1 score is important)
- To minimize rejection of non-defaulters, in order to reduce loss by incorrect denials (class 0 Precision is also important)
- To have accuracy greater than 80%, since about 20% of applications are defaulters with current approval process
- To identify important independent features that help in classification of loan applicants as defaulters or non-defaulters based on the predictive model.

Challenges in Data

- This data has imbalanced target classes, as 80% of loans are PAID (indicated by class 0), whereas only 20% are DEFAULTED (class 1). Performance tuning techniques that can treat this imbalance of target classes were used, which helped in improving model performances to a large extent.
- Preliminary data exploration reveals data quality issues like outliers, skewed distribution of independent variables, high percentage of missing data for certain features, that can affect model development. These are taken into consideration while data cleansing, data pre-processing and feature engineering steps.
- Some of the variables seem to have high multicollinearity. This could interfere in the performance of the logistic regression model. Thus, multicollinearity was checked for using VIF score.

Solution design approach

Since this is a Binary Classification problem, with imbalanced data, our focus will be on following techniques:

- Data preparation techniques:
  - Separate Independent and Dependent variables
  - Split data into training and test sets using stratification on target variable
  - Check for multicollinearity
  - Treat data for imbalance using SMOTE technique for Logistic regression model
- Classification techniques:
  - Logistic Regression - baseline, optimized, resampled data
  - Decision Tree - baseline, tuned
  - Random Forest - baseline, tuned
- Model Analysis techniques:
  - Confusion Matrix
  - Classification Report
  - Feature importances

Performance tuning techniques explored to achieve balance of classes in data, to improve models performances, and avoid overfitting problems are:

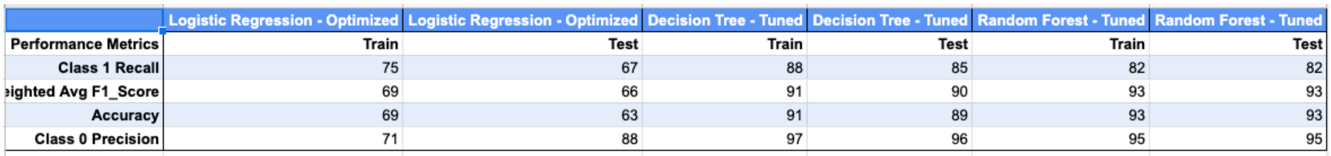
1) Stratify train-test dataset splitting over the y-variable (binary stratification) - this helped with increased Class 1 Recall and improved metrics for LR model.
2) Optimal threshold of Precision-Recall curve for tuning Logistic Regression model - helped in enhanced metrics for baseline LR model
3) Resampling of training data using SMOTE - had a potential to enhance metrics by oversampling of minority class, but did not deliver better metrics as compared to first two techniques
4) Class-weights in decision tree and random forest models - using inverted weights for the imbalanced classes, we achieved a balanced (1:1) proportion of classes. It helps with purity of nodes to get unbiased classification
5) GridsearchCV using 5-fold Cross-validation for hyper-parameter tuning of Decision Tree model and Random Forest model
6) Feature Engineering to include new logical feature ‘STATUS’ and to transform discrete categorical features ‘REASON’, ‘JOB’, and continuous categorical features “DEROG’ and ‘DELINQ’.

Performance Metrics used to analyze models at each stage of development are:

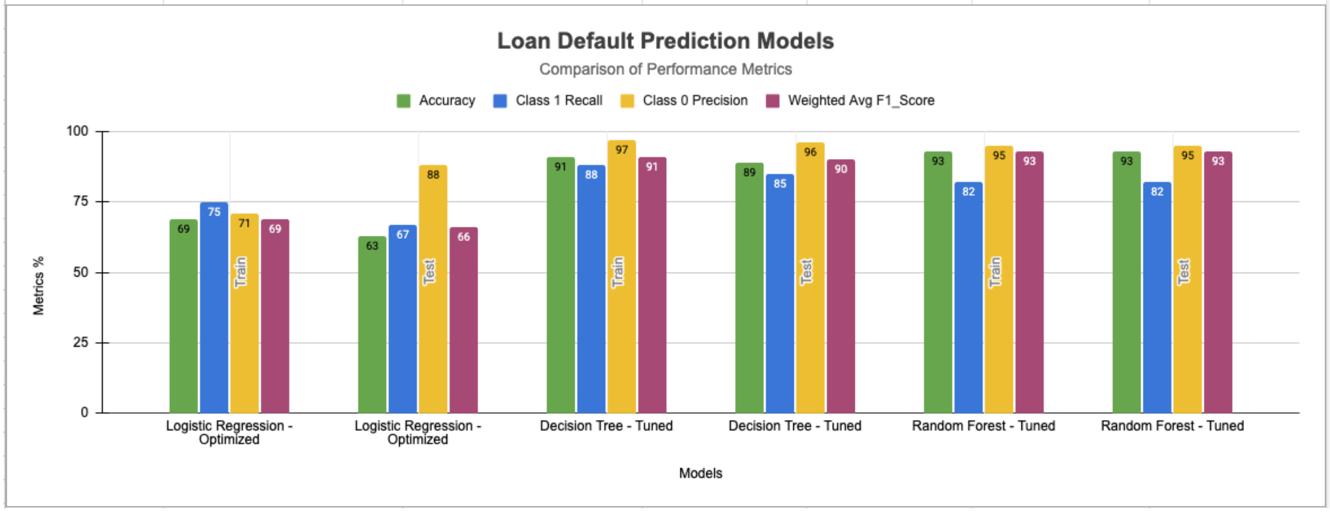
- Recall Score for Class 1 (Retrieval of most relevant records from data)
- Weighted Average F1 Score (Achieve balance of classes 0 and 1)
- Weighted Average Precision (Balanced accuracy of retrieved records)
- Weighted Average Recall Score (Balanced retrieval of relevant records)
- Accuracy of the Model (Accuracy of Prediction)

Model comparison

Below is a detailed overview of model performance metrics for best version of each type of classification technique used.



	Logistic Regression - Optimized	Logistic Regression - Optimized	Decision Tree - Tuned	Decision Tree - Tuned	Random Forest - Tuned	Random Forest - Tuned
Performance Metrics	Train	Test	Train	Test	Train	Test
Class 1 Recall	75	67	88	85	82	82
eighted Avg F1_Score	69	66	91	90	93	93
Accuracy	69	63	91	89	93	93
Class 0 Precision	71	88	97	96	95	95



Since regression techniques are not robust to outliers, the data used for this model was treated for outliers, and missing values were replaced using mean values.

Logistic Regression Model

- Baseline model has poor score for predicting class 1, with model accuracy of 81%
- Performance tuning using optimal threshold improves the class 1 Recall score, but lowers accuracy
- Model fitted on resampled data using SMOTE technique has best scores such as 67% class 1 recall, 66% weighted average F1_score, 63% accuracy

Since tree based models techniques are robust to outliers, the data used for this model was NOT treated for outliers, but missing values were replaced using mean values

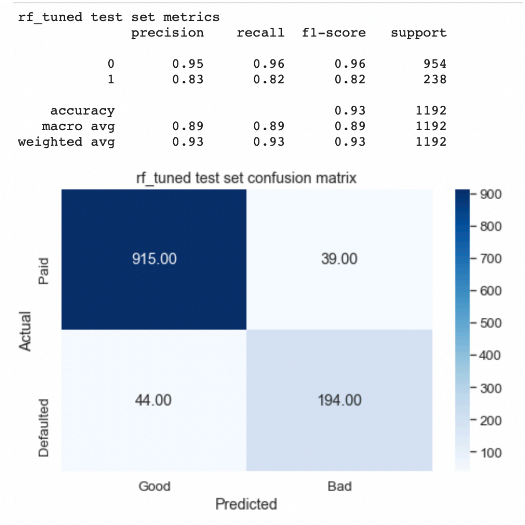
Decision Tree Model

- Baseline model is overfitted on training data, but has good generalization with 95% accuracy on test data
- After hyper-parameter tuning, the model has a good fit on training as well as test data
- Tuned decision tree model has 85% recall rate for test data, with 89% accuracy

Regression Forest Model

- Baseline model is overfitted on training data, but has good generalization with 95% accuracy on test data
- After hyper-parameter tuning, the model has a very good fit on training as well as test data

Final Proposed Solution

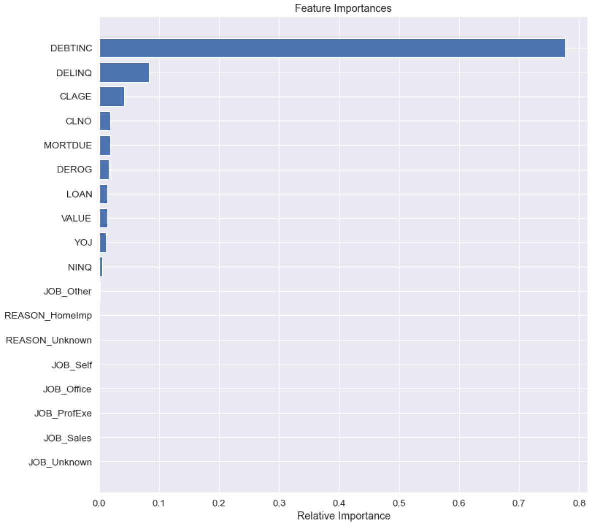


```
rf_tuned test set metrics
              precision    recall  f1-score   support

           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
```

- Tuned Random Forest model has a very good 82% recall rate on train and test data
- 93% accuracy which is highest of all models
- Tuned RF is the best model considering a good class 1 recall score, combined with high class 0 precision of 95%
- It has weighted average F1_score of 93%, which indicates that the model will have a good balance between identifying most probable Defaulters as well as non-defaulters.



- Most important features of the tuned RF model are DEBTINC, DELINQ, CLAGE, CLNO,followed by MORTDUE, DEROG, LOAN, VALUE, YOJ and NINQ. We had anticipated most of these features to be predictor variables, which is highlighted with this result.
- Thus, the tuned Random Forest model is the best model considering it minimizes the errors and biases of manual process and gives more accurate predictions in much lesser time.

Recommendations

Benefits of proposed solution

- As discussed in the executive summary, there is a significant reduction in loss of principal amount based on accuracy of prediction of defaulters. Estimated reduction in this loss is 75%, which can far outweigh the loss of losing some good loans in anticipation of not approving probable defaulters. Business could review the cost benefits of replacing existing approval process with the proposed solution on the 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 and faster processing times

Data considerations

- From a development perspective, the machine learning classification algorithms work best if the data has maximum valuable information in a substantial quantity. For e.g., in our case, this dataset has less than 10,000 observations, with missing data in the range of 4-21% in almost 90% of the features. The already stochastic nature of machine learning algorithms is further extrapolated with such factors. Thus, **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.
- Better insights into the consumer's credit profile can help with developing deeper predictive models with enhanced performance. Adding more features to source data such as credit score, income range, interest rates can be beneficial for such analysis.

Technical considerations

- We can further explore treating outliers techniques such as PowerTransformer, QuantileTransformer, or Winsorizing.
- Missing data imputation can be tried using SimpleImputer.
- Pipeline methods of model building can be explored.
- Performance tuning techniques could need revision if there is major change in data quality in the future
- Deployment techniques for the further application of this proposed model will need to be explored based on business preferences

Business Considerations

- Legal challenges with machine learning model governance and review need to be considered before deployment in production
- Solution design can be further enhanced to suggest approval for partial loan amounts, or loan at interest rates in proportion to the credit risk of the loan applicant.