# Overview

## Introduction

The following is a predictive analysis to determine the eligibility of loan applications based on credit history.

## Objective

Predict ‘Bad\_flag\_worst6’ target feature with a model built on training set, then evaluating the model on the given test set.

This is a binary classification problem, with bad\_worst\_flag 6 having classes ‘0’ and ‘1’, denoting ‘accept’ and ‘reject’ loan application response respectively.

As this problem entails the analysis of loans, some metrics are of special importance must be considered. They include the following:

* **Repayment History**  – the punctuality of loan and credit card payment dues.
* **Credit Utilization** – The ratio of total credit (liabilities) is owed to total credit limit. A very high credit utilization ratio is considered risky.
* **Duration of Credit Servicing** – The amount of time for which loans have been serviced. A lengthy history with timely payments would reflect positively.
* **Credit Mix** – The more widespread the mix of credit (spread across credit cards, auto loan, personal loan, home loan etc.), the better.
* **Number of Credit Enquiries** – Indicates how frequency of loan applications. Credit-hungry applicants are riskier.

# Data Pre-Processing and Analysis

Pre-processing was done remove duplicate samples, correct erroneous data, set features to the correct data type and remove features that contribute minimally to the predictive power of the model to be built; e.g. feature with all NA values, features with a single value throughout, features containing information from other variables, and features that are irrelevant to the prediction. Both the train and test sets were cleaned simultaneously, but only the train set will be used for exploratory analysis since it is meant to be unseen data.

The predictor variables were analysed using R’s ‘FSelector’ package, which calculates the weight of the importance of the attribute with respect to the target variable using information gain.

**information gain = H(class) + H(attribute) - H(class,attribute)**

where H(class) is entropy of class and H(class, attribute) is conditional entropy of class given attribute.

The training set contains information for 23896 unique applications for loans and it is split into three subsets named ‘Data’, ‘Account’ and ‘Enquiry’.

## ‘Data’

**23896** observations

Initially containing 87 features of current information for the loan applications, most importantly the response feature ‘bad\_worst\_flag6.’ It mostly consists of demographic information like gender, place of residence, occupation, email address, etc.

### Target Feature

|  |  |  |
| --- | --- | --- |
| **bad\_flag\_worst6** | | |
| **Class** | **0** | **1** |
| **Count** | 22892 | 1004 |
| **Percentage(%)** | 95.8 | 4.2 |

The target feature ‘bad\_worst\_flag6’ is imbalanced - of the 23896 samples, only 1004 are of class ‘1’. As this is a classification problem, this will skew the predictive model to favour predictions of the ‘0’ class significantly.

Measures like oversampling and undersampling - where the minority class is resampled and the majority class is undersampled respectively - will be required to rectify the imbalance.

### Predictor Features

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature Importance Based on Information Gain** | | | | |
| **Feature** | **worst\_dpd6** | **designation** | **office\_pin** | **app\_res\_pincode** |
| **Weight** | 0.17429714 | 0.05171279 | 0.01441458 | 0.01423210 |
|  |  |  |  |  |
| **Feature** | **se\_code** | **cibil\_datetime** | **acq\_source** | **existing\_card\_start\_date** |
| **Weight** | 0.00923756 | 0.00632221 | 0.00602530 | 0.00436457 |
|  |  |  |  |  |
| **Feature** | **promo\_code** | **approved\_credit\_limit** | **lead\_code** | **existing\_credit\_limit** |
| **Weight** | 0.00350787 | 0.00319897 | 0.00293240 | 0.00218101 |
|  |  |  |  |  |
| **Feature** | **existing\_card\_issuer** | **app\_has\_card** | **fee\_code** | **app\_res\_city** |
| **Weight** | 0.00200780 | 0.00184053 | 0.00162837 | 0.00151035 |
|  |  |  |  |  |
| **Feature** | **office\_city** | **cibil\_score** | **existing\_bank** | **card\_name** |
| **Weight** | 0.00143775 | 0.00122086 | 0.00118932 | 0.00095652 |
|  |  |  |  |  |
| **Feature** | **app\_dob** | **override\_fee\_code** | **industry\_type** | **aip\_status** |
| **Weight** | 0.00060229 | 0.00052293 | 0.00041372 | 0.00032270 |
|  |  |  |  |  |
| **Feature** | **reject\_reason\_code** | **reject\_reason\_desc** | **company\_type** | **marital\_status** |
| **Weight** | 0.00031023 | 0.00031023 | 0.00025089 | 0.00023443 |
|  |  |  |  |  |
| **Feature** | **edu\_qualification** | **mob\_verified** | **employment\_type** | **intl\_trn** |
| **Weight** | 0.00022253 | 0.00017265 | 0.00014784 | 0.00006893 |
|  |  |  |  |  |
| **Feature** | **mktg\_code** | **res\_type** | **app\_gender** | **permanent\_same** |
| **Weight** | 0.00005848 | 0.00003685 | 0.00001734 | 0.00001123 |
|  |  |  |  |  |
| **Feature** | **net\_monthly\_income** | **dt\_opened** | **override\_months** | **num\_dependents** |
| **Weight** | 0.00000405 | 0.00000000 | 0.00000000 | 0.00000000 |
|  |  |  |  |  |
| **Feature** | **res\_from\_yr** | **year\_joining** | **years\_exp** |  |
| **Weight** | 0.00000000 | 0.00000000 | 0.00000000 |  |

The ‘Data’ subset consists mostly demographic (personal particulars) data and cannot be used to predict viability of loan disbursement, even if there seems to be a relatively moderate information gain.

The ‘worst\_dpd6’ feature containing the worst days past due values proves to be most useful, but in actuality it is a feature that directly defines the target feature and as such is unusable. Other features include ‘approved\_credit\_limit’, ‘existing\_credit\_limit’, and ‘cibil\_score’ as they relate to financial measures.

## ‘Account’

**186039** observations

### Predictor Features

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature Importance based on Information Gain** | | | | |
| **Feature** | **paymenthistory1** | **paymenthistory2** | **opened\_dt** | **acct\_type** |
| **Weight** | 0.02200695 | 0.01433556 | 0.00074610 | 0.00068117 |
|  |  |  |  |  |
| **Feature** | **paymt\_end\_dt** | **cur\_balance\_amt** | **creditlimit** | **reporting\_dt** |
| **Weight** | 0.00052823 | 0.00031899 | 0.00026638 | 0.00024453 |
|  |  |  |  |  |
| **Feature** | **paymt\_str\_dt** | **cashlimit** | **high\_credit\_amt** | **writtenoffandsettled** |
| **Weight** | 0.00023205 | 0.00017991 | 0.00016666 | 0.00014378 |
|  |  |  |  |  |
| **Feature** | **last\_paymt\_dt** | **paymentfrequency** | **typeofcollateral** | **closed\_dt** |
| **Weight** | 0.00010024 | 0.00005531 | 0.00004308 | 0.00002654 |
|  |  |  |  |  |
| **Feature** | **owner\_indic** | **cibilremarkscode** | **amt\_past\_due** | **valueofcollateral** |
| **Weight** | 0.00000499 | 0.00000254 | 0.00000000 | 0.00000000 |
|  |  |  |  |  |
| **Feature** | **rateofinterest** | **repaymenttenure** | **writtenoffamounttotal** | **writtenoffamountprincipal** |
| **Weight** | 0.00000000 | 0.00000000 | 0.00000000 | 0.00000000 |
|  |  |  |  |  |
| **Feature** | **settlementamount** | **actualpaymentamount** |  |  |
| **Weight** | 0.00000000 | 0.00000000 |  |  |

The ‘Account’ subset consists of the most important data, mainly of the applicant’s financial background, mostly relating to loans and credit payment.

Features describing payment history seem to have the most significance. However they are stored as a series of characters and further processing must be done to interpret it better.

Various other features will be of higher importance than the rest- ‘creditlimit’, ‘cur\_balance\_amt’, ‘cashlimit’, ‘high\_credit\_amt’, ‘paymentfrequency’.

## ‘Enquiry’

**404034** observations

### Predictor Features

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature Importance based on Information Gain** | | | | |
| **Feature** | **dt\_opened** | **enquiry\_dt** | **enq\_purpose** | **enq\_amt** |
| **Weight** | 0.006993 | 0.001272 | 0.000761 | 0.000329 |

The ‘Enquiry’ subset contains information regarding credit enquiries. Information pertaining to the frequency, recency and type of loan enquiry would be useful.

# Feature engineering

Features relating to repayment history, credit utilization, duration of credit servicing, credit mix and credit enquiries are good metrics for determining loan risk.

The following are features that were constructed/selected and tested using a Random Forest model.

## Features

### payment\_history\_avg\_dpd\_0\_29\_bucket

Average of the count of each payment falling into the less-than-30-days-past-due bag, with less than 30 days past due being a ‘punctual’ payment and therefore lower risk.

### payment\_history\_avg\_dpd\_30\_bucket

Average of the count of each payment falling into the 30-or-more-days-past-due bag, with more than 30 days past due being a ‘late’ payment and therefore higher risk.

### payment\_history\_mean\_length

Mean length of the payment history, indicating duration of loan payments.

### pay\_dpd\_ratio

Ratio of the average number of payments falling into the 30-or-more-days-past-due bag against average number of payments. Indicates duration of non-payment throughout payment history.

### total\_diff\_lastpaymt\_opened\_dt

The total duration between last payment date and account opened date of all accounts. Shows recency of payment.

### min\_months\_last\_30\_plus

Minimum number of months that passed before first 30-or-more-days-past-due bag appeared, showing punctuality.

### utilisation\_trend

[total cur\_bal\_amt / total credit limit] / [mean cur\_bal\_amt / (mean credit limit+ mean\_cashlimit)]

### ratio\_currbalance\_creditlimit

Ratio of current balance to credit limit, showing how much applicant is dependent on credit.

### count\_enquiry\_recency\_365

Number of enquiries made in the last 365 days. Shows frequency of loan application

### count\_enquiry\_recency\_90

Number of enquiries made in the last 90 days. Shows frequency of loan application

### mean\_diff\_open\_enquiry\_dt

Average difference between enquiry dt\_opened date and enquiry date. Shows recency of the previous loan application.

### enq\_purpose

Most frequent enquiry purpose. Unsecured loans would be riskier.

### worst\_dpd

Worst number of days past due. Shows ability to pay dues.

### cibil\_score

Raw CIBIL score.

### cibil\_bag

Bagged CIBIL score.

### approved\_credit\_limit

As given in raw data.

### existing\_credit\_limit

As given in raw data.

# Model Evaluation

## Model Features

A Random Forest model was built with the following features, ranked by their importance. Over- and under-sampling were used to balance the number of observations for each class of the target feature.

|  |  |
| --- | --- |
| **Feature Importance based on Information Gain** | |
| **Feature** | **Weight** |
| **pay\_dpd\_30\_bucket** | 0.018357 |
| **approved\_credit\_limit** | 0.016087 |
| **payment\_history\_mean\_length** | 0.014856 |
| **pay\_dpd\_0\_29\_bucket** | 0.012390 |
| **ratio\_currbalance\_creditlimit** | 0.008537 |
| **existing\_credit\_limit** | 0.008284 |
| **mean\_diff\_open\_enquiry\_dt** | 0.006787 |
| **pay\_dpd\_ratio** | 0.006201 |
| **total\_diff\_lastpaymt\_opened\_dt** | 0.006048 |
| **enq\_purpose** | 0.004169 |
| **cibil\_bag** | 0.003913 |
| **worst\_dpd** | 0.000000 |
| **min\_months\_last\_30\_plus** | 0.000000 |
| **utilisation\_trend** | 0.000000 |
| **count\_enquiry\_recency\_365** | 0.000000 |
| **count\_enquiry\_recency\_90** | 0.000000 |
| **cibil\_score** | 0.000000 |

## Model Evaluation

### GINI

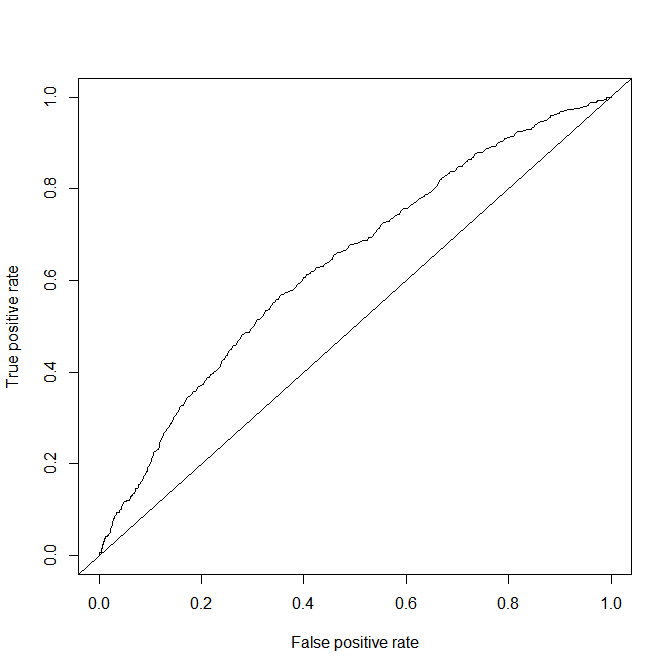
GINI was calculated as Gini = 2 \* Area Under ROC – 1

Max. GINI: 27.26

### Rank Ordering

|  |  |
| --- | --- |
| **Decile** | **True Positive** |
| **10** | 0.00% |
| **9** | 0.88% |
| **8** | 1.08% |
| **7** | 2.64% |
| **6** | 6.65% |
| **5** | 7.33% |
| **4** | 15.15% |
| **3** | 24.14% |
| **2** | 32.94% |
| **1** | 41.05% |

### ROC



### Specificity, Sensitivity

Specificity: 0.8682567

Sensitivity: 0.1341991

# Summary

ROC and GINI indicate that the predictive power of the model is somewhat viable, however rank ordering and sensitivity highlights that most of the true positives are in the lower decile, with low probability of prediction. A sensitivity of 0.13 means that the model only captures 13% of the risky loans. This is partially due to the severely imbalanced data set with limited positive observations.

Although this model can achieve a maximum GINI of 27.26, adjustments may be done to improve sensitivity to 0.25, at the cost of reducing specificity to 0.73. Since risk prediction leans towards risk mitigation, a higher sensitivity may be preferred at the cost of lower specificity (i.e. rejecting less risky loans).

Future improvements to this model will need to include better features that capture the essence of the risky loans better, despite the rarity of positive samples, perhaps with cost-sensitive learning.