MACHINE LEARNING APPLICATION IN LOAN DEFAULT PREDICTION

ABHISHEK KUMAR TIWARI
Manager, Tata Consultancy Services
tiwariabhishek1120@gmail.com

ABSTRACT:

In Todays world, most of world population has access to banking services. Consumers has increased many fold in last few years. For the banks, risks related to bank loans has increased especially after The Great Recession (2007–2012) and job threats due to automation and advancement in technologies like artificial intelligence (AI). At the same time technological advancement enabled companies to gather and save huge data which represent the customer's behavior and the risks around loan.Data Mining is a promising area of data analysis which aims to extract useful knowledge from tremendous amount of complex data sets

Non-Performing Assets (NPA) is the top most concerns of banks. The NPA list is topped by PIIGS (Portugal, Italy, Ireland, Greece and Spain) countries

Introduction

This paper proposes the use of statistical methods especially machine learning techniques to model and predict bank losses. We have used different machine learning algorithms specifically designed to handle computationally intensive recognition of interaction in large data-sets. The methods use all information available regardless on prior beliefs about their importance, and take into account of their interaction effects among all variables (Features). We have applied four machine learning algorithms to predict Loan Default Prediction: Logistic Regression, K-Nearest Neighbors (KNN), the tree-based classifier, Classification and Regression Tree (CART) and Random Forest (RF). These models are suited for Loan Default Prediction because of the large sample sizes and complexity of the possible relationships among variables. The data is split into a training data-set (75%) for model development and a testing data-set (25%) used for out of sample prediction.

The machine learning methods used have their pros and cons.

MACHINE LEARNING ALGORITHMS

LOGISTIC REGRESSION:

The logistic regression is the most widely used techniques for classification purpose. It expresses the linear regression equation in logarithmic terms, called the logit or log of odd where

odd = probability of success / probability of failure

Response variable for this paper is "**Default**", where the default has binary outcome, Yes or No. Logistic regression models the probability of success in this case (probability of being Default, Default = Yes). Logistic Regression is defined by sigmoide function, a S-shaped curve shown below

$$\log(Y) = \log(Y/(1-Y)) = \beta 0 + \sum_{i=1}^{p} Xi\beta i$$

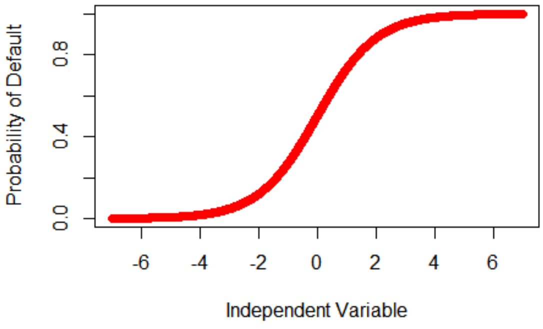


Fig 1: Sigmoid Function

K-NEAREST NEIGHBORS (KNN)CLASSIFIER:

K-nearest neighbor classification method, a very simple method that works really well on many problems/dataset. KNN classifier first calculate the distance and identifies the K points (neighbors) in the training data that are closest to x_0, represented by N_0, given a positive integer K and a test observation x_0 .It then estimates the conditional probability for class j as the fraction of points in N_0 whose response values equal j

$$\Pr(Y = j \mid X = x0) = \frac{1}{K}\sum_{i\in No} I(yi = j)$$

And then finally, KNN classifies the test observation x_0 to the class with the largest probability by applying Bayes rule.

CLASSIFICATION AND REGRESSION TRESS (CART):

CART uses Gini Index as The impurity (or purity) measure used in building decision tree

$$Gini = \sum_{i\neq j} p(i)p(j)$$

Where i and j are levels of target variable

Minsplit and Minbucket is important parameter in CART. **minsplit** is minimum number of observation for split attempt, **minbucket** is minimum number of observation in leaf node. CART can work even on unbalanced data by changing the prior probabilities to obtain a decision tree

parms = list(prior=c(non_default_proportion, default_proportion))

or by including a loss matrix as

parms = list(loss = matrix(c(0, cost_def_as_nondef, cost_nondef_as_def, 0), ncol=2))

Detailed explanation of these methods are explained in imbalanced data section.

Trees obtained by CART are easily interpretable and very easy to explain, very useful in case we want to translate rules in English. Less data preparation is required. CART is very robust to outliers in the input variables. CART can use the same variables multiple times in different parts of the tree. This can uncover complex interactions between sets of variables.

Final estimate of tree can change with a small change in the data. CART usually overfits which can be solved by pruning of tree.

RANDOM FOREST:

Random forests, use trees as building blocks to construct more powerful prediction models. It improves accuracy by fitting many trees by small tweak that decorrelates the trees. A random sample of m variables is chosen as split candidates from the full set of p variables. mtry is number of randomly selected variables used at each split

$$m = \sqrt{p} \text{ for classification and } m = \frac{p}{3}$$

for Regression

For each bootstrap sample it grow un-pruned tree by using best split based on mtry at each node

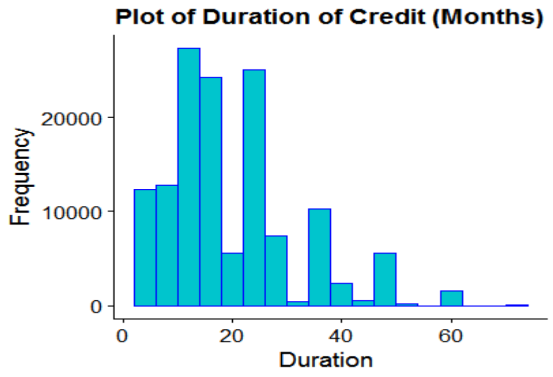
Random Forest predict the test data by choosing majority class for Classification and by taking mean in case of Regression. Random forests have a tendency to bias towards variables that have more number of distinct values i.e. favor numeric variables over binary/categorical values.

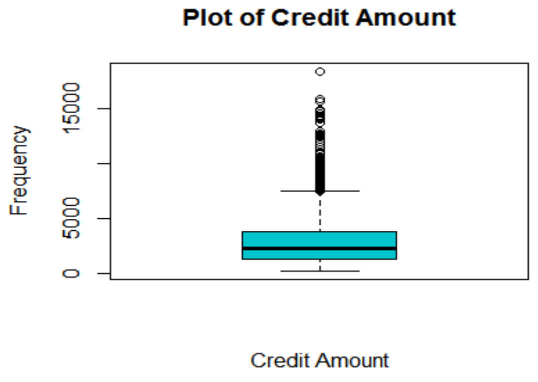
MODEL BUILDING METHODOLOGY:

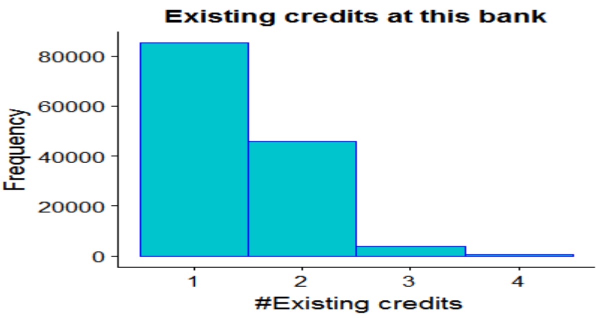
Steps involved in this model building methodology are mentioned below:

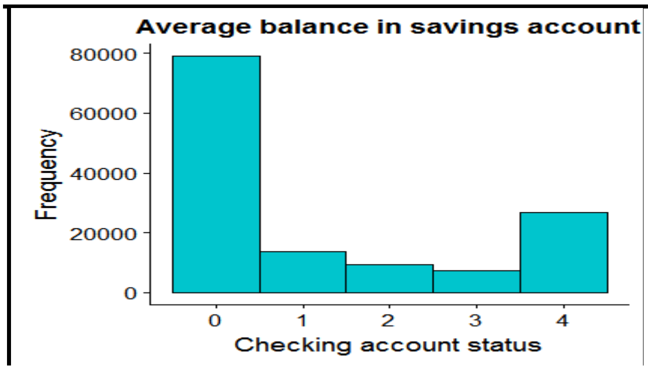
- Data Selection
- Exploratory Data Analysis
- Outlier Detection
  - Outlier Treatment
- Missing Value Detection
  - Missing Value Treatment
- Splitting Training & Test Datasets
- Check for Data Imbalance
  - Over Sampling
  - Under Sampling
  - SMOTE
  - Changing the prior probabilities
  - Loss matrix
- Features Selection
- Building Classification Model
- Predicting Class Labels of Test Dataset
- Evaluating Accuracy and other metrics
- Parameter Tuning
- Finalize the Model

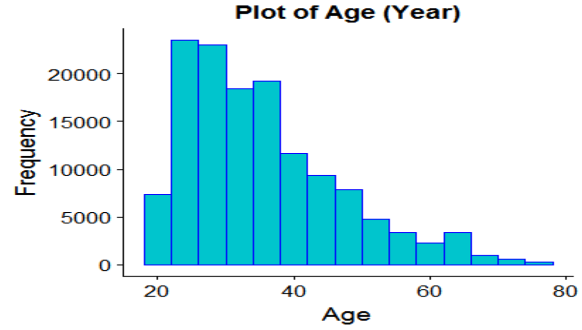
EDA (EXPLORATORY DATA ANALYSIS):

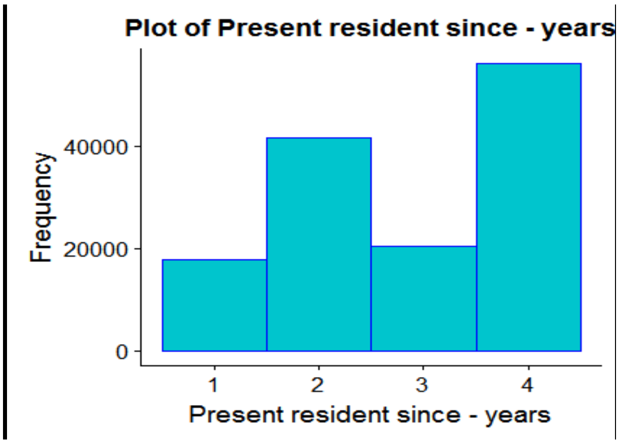


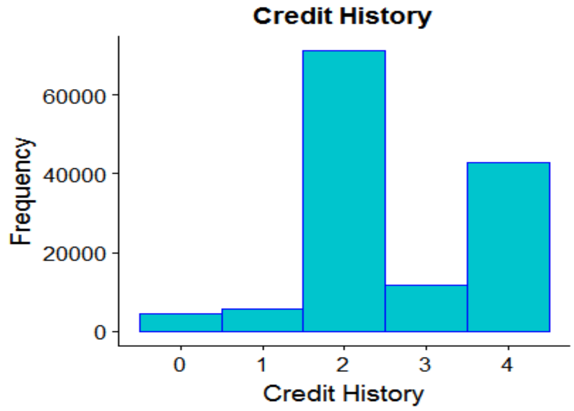












OUTLIERS DETECTION AND TREATMENT:

Any observations which is not in range Q1 - 1.5*IQ to Q3+ 1.5*IQ can be considered as outliers. We need to be sure before treating/eliminating outliers that it's not influential observation.

Graphical method:

Box Plot

outlier_cutoff_high<- quantile(Data$Var, 0.75) + 1.5 * IQR(Data$Var)

outlier_cutoff_low<- quantile(Data$Var, 0.25) - 1.5 * IQR(Data$Var)

Replace observations beyond higher cutoff point by 95th percentile and lower cut off by 5th percentile

Statistical Technique

Grubb's test for outliers

R package "Outlier"

MISSING VALUE TREATMENT

Drop variable if it has more than 30% of missing values Drop Observation if has many attributes missing

IMPUTATION OF MISSING VALUES:

Multivariate Imputation by Chained Equations (MICE) Very powerful and popular technique for imputation of missing values, it uses different default methods for different kind of data

Numeric data :: **pmm**, predictive mean matching

Binary data with 2 levels :**logreg**, logistic regression imputation

Unordered categorical data (factor >= 2 levels): **polyreg**, polytomous regression imputation

Orderedcategorical data (factor>= 2 levels)**polr**, proportional odds model

KNN IMPUTATION:

Impute with neighbor based on existing attributes by using Euclidean or Manhattan distance

DATA IMBALANCE:

Oversampling methods replicate the observations from the minority class tobalance the data, this may cause overfiting.

Under-sampling methods remove the majority of classes to balance data. Removing observations causes the loss of useful information pertaining to the majority class.

Synthetic Minority Oversampling Technique (SMOTE) finds random points within nearest neighbors of each minor class observation and by boosting methods generates new minor class observations. New data are not the same as the existing data it does not have any overfiting problem

Changing the prior probabilities: Changing the prior probabilities to obtain a decision tree, This is an indirect way of adjusting the importance of mis-classifications for each class.

parms = list(prior=c(non_default_proportion, default_proportion))

Including a loss matrix: Loss matrix can be included, changing the relative importance of misclassifying a default as non-default versus a non-default as a default.

Ifproblem demands that misclassifying a default as a non-default should be penalized more heavily. Including a loss matrix can again be done in the argument parms in the loss matrix.

parms = list(loss = matrix(c(0, cost_def_as_nondef, cost_nondef_as_def, 0), ncol=2))

Doing this, we are constructing a 2x2-matrix with zeroes on the diagonal and changed loss penalties off-diagonal. The default loss matrix is all ones off-diagonal

Features Selection

Graphical representation of the variable importance for Top 10 variablemean decrease in Gini index) is shown in figure. The variables with the largest mean decrease in Giniindex are Amount, Chk_Acct, and Duration. Gini index is decreased by splits over a given predictor, averaged over all trees. Large number indicates high importance.

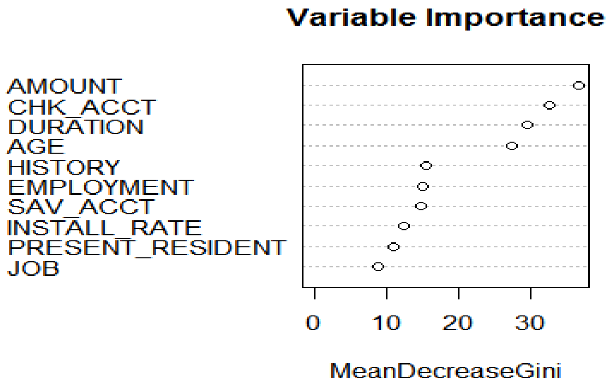
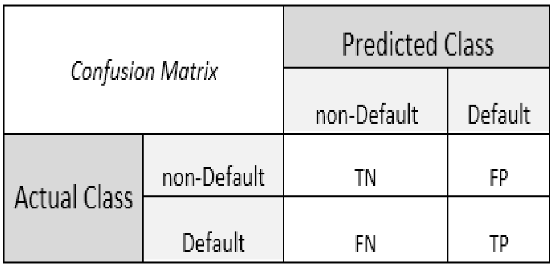


Fig 2: Variable Importance plot

PREDICTION ACCURACY METRICS:

CONFUSION MATRIX:



Confusion Matrix		Predicted Class	
		non-Default	Default
Actual Class	non-Default	TN	FP
	Default	FN	TP

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$\Pr ecision = \frac{TP}{TP + FP}$$

F1 scoreis the harmonic
mean of precision and sensitivity (recall)

$$F1 = \frac{2TP}{2TP + FP + FN}$$

RECEIVER OPERATING CHARACTERISTICS (ROC):

The ROC curve is a popular graphic which simultaneously display the two types of errors for all possible thresholds, the vertical axes is the true positive rate (Sensitivity) and the horizontal axes is the false positive rate (1-Specificity) for different threshold points of a parameters. If the curve is closer to the top left then the accuracy of the prediction is higher. ROC curves are useful for comparing different classification algorithm.

According to the ROC curve, Random Forest and KNN models have the highest accuracy and the CART models have the lowest.

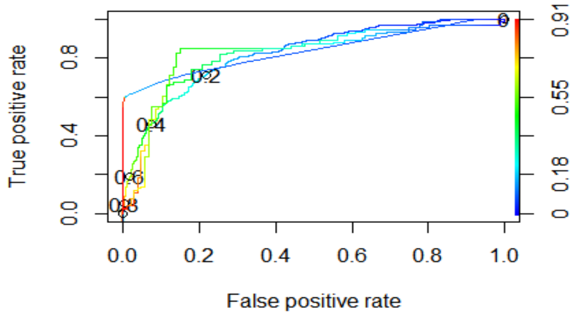


Fig 3: ROC curve for different classifiers

AREA UNDER THE CURVE (AUC):

AUC is a metric for binary classification that measures the accuracy of model, ranging from 0.5 to 1

PARAMETER TUNING :

CP (Complexity Parameter)

Misclassification rate = Root node error * Xerror * 100%

Pick the Cp value which corresponds to least Xerror and use this Cp to prune the treee.

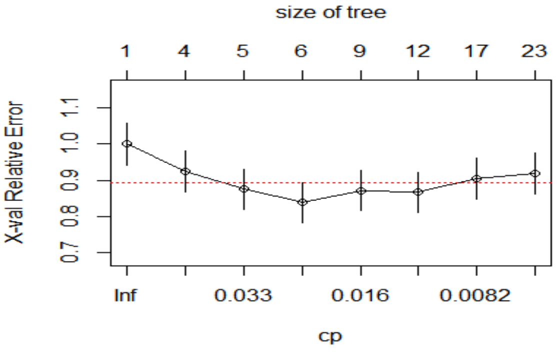


Fig 4: Cp value VsXerror

N TREE (NUMBER OF TRESS IN RANDOM FOREST):
Choose the number of trees from plot where elbow is formed, error does not decrease significantly. Random Forest with lesser number of tress will be faster in execution.

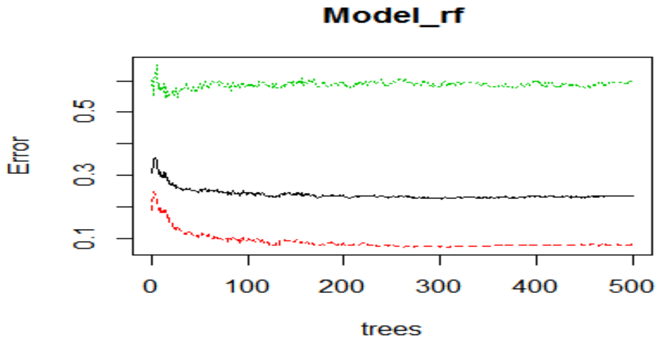


Fig 5: Number of trees Vs error

MTRY
Number of variables randomly sampled at each split

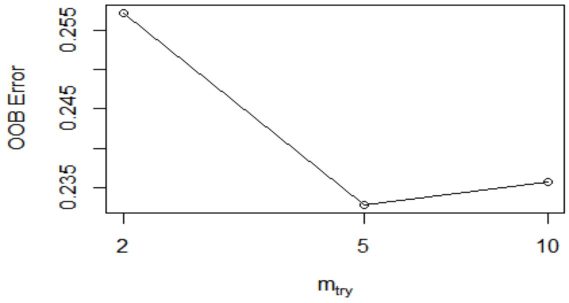


Fig 6: mtryVs OOB error

CONCLUSION:

Accurate Default estimation can help banks to avoid huge losses. In this paper we presented a framework for effectively prediction the class labels of the new loan applicants. These model were built using the data mining techniques available in the R. Preprocessing step is the most important and time consuming part. Pre-processed dataset is then used for building the decision tree classifier.The results, verify that machine learning algorithms yield higher forecast accuracy. Machine learning algorithms can help to recognize the importance of the variables.The results indicate that these four machine learning methods has its pros and cons. We evaluate the prediction performance using the metric Area Under the Curve (AUC), F1 score, Recall, Precision, Accuracy using the ROC Curve,which plots the true positive rates against false positive rates. According to these metrics, Random Forest and KNN models have the high accuracy and the CART models have the lowest. Random Forest has given 86% accurate classification result.

REFERENCES:

1) Gareth James, Trevor Hastie, Daniela Witten and Robert Tibshirani, “An Introduction to Statistical Learning” E.N. Hamid, and N. Ahmad, “A New Approach for Labeling the Class of Bank Credit Customers via Classification Method in Data Mining”, International Journal of Information and Education Technology, vol. 1(2), pp. 150-155, 2011.
2) K. Kavitha, “Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques”, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6(2), pp. 162–166, 2016.
3) Z. Somayyeh, and M. Abdolkarim, “Natural Customer Ranking of Banks in Terms of Credit Risk by Using Data Mining A Case Study: Branches of Mellat Bank of Iran”, Jurnal UMP Social Sciences and Technology Management, vol. 3(2), pp. 307–316, 2015.
4) M. Sudhakar, and C.V.K. Reddy, “Two Step Credit Risk Assessment Model For Retail Bank Loan Applications Using Decision Tree Data Mining Techniq