**Statistical Modelling for Small and Medium Enterprise Credit Lending in India**

***A Thesis***

***Submitted in the partial fulfilment of the***

***requirements for the award of the Degree of***

**INTEGRATED MASTER OF SCEINCE**

**IN**

**MATHEMATICS AND COMPUTING**

BY

**ARPIT KUMAR**

**(IMH/10011/2012)**

****

**DEPARTMENT OF MATHEMATICS**

**BIRLA INSTITUTE OF TECHNOLOGY**

**MESRA-835215, RANCHI**

**DECLARATION**

I certify that

a. The work contained in the thesis is original and has been done by myself under the general supervision of my supervisors.

b. The work has not been submitted to any other institute for any degree or diploma.

c. I have followed the guidelines provided by the Institute in writing the thesis.

d. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

e. Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

f. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Name: Arpit Kumar

Roll No.: IMH/10011/2012

**APPROVAL OF THE GUIDE**

Recommended that the thesis entitled **“Statistical Modelling for Small and Medium Enterprise Credit Lending in India”** prepared by Arpit Kumar under my supervision and guidance be accepted as fulfilling this part of the requirements for the degree of Integrated Master of Science in Mathematics & Computing.

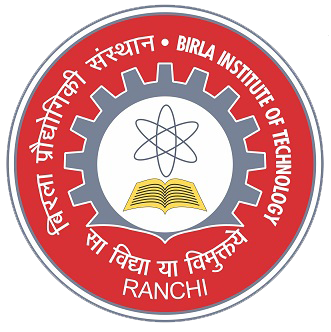
To the best of my/our knowledge, the contents of this thesis did not form a basis for the award of any previous degree to anyone.

Date:

Dr. Saral Kumar Jain

Department of Mathematics

Birla Institute of Technology, Mesra



**THESIS APPROVAL CERTIFICATE**

This is to certify that the work embodied in this thesis entitled “**Statistical Modelling for Small and Medium Enterprise Credit Lending in India**” carried out by Mr. Arpit Kumar (Roll No.: IMH/10011/2012) is approved for the degree of Integrated Master of Science in Mathematics and Computing of Birla Institute of Technology, Mesra, Ranchi.

Date:

Place:

Internal Examiner(s) External Examiner(s)

Name & Signature Name & Signature

Chairman

(Head of the Department)

**ACKNOWLEDGEMENT**

I take immense pleasure in thanking Dr. S.K. Jain, Department of Mathematics, Birla Institute of Technology, Mesra for having me carry out this internship and for his constant guidance & support.

I would like to express my heartfelt gratitude to Capital Float and Mr. Tushar Garimalla, Vice President, Decision Sciences, for his relentless support and encouragement throughout the course of the project. This project would not have been possible without his guidance, and I shall forever remain indebted to him. for giving me the opportunity to work under their guidance and help me gain immensely enriching professional experience.

Arpit- Kumar, IMH/10011/2012

Department of Mathematics

BIT Mesra

# Abstract

Thesis Title: Statistical Modelling for Small and Medium Enterprise Credit Lending in India

|  |  |  |
| --- | --- | --- |
| Supervisor: | Mr. Tushar Garimalla |  |
| Semester: | Tenth | Year: 2016-17 |
| Name of the student: | Arpit Kumar | Roll No.: IMH/10011/12 |

Abstract:

*Small and Medium Enterprises are seen as the engines of growth for India today. They account for a large proportion of India’s manufacturing output and contribute significantly towards the nation’s GDP. One of the major hindrances that SMEs face is the lack of quick and easy access to credit. The need for collateral, excessive documentation and rigid terms has led to SMEs being unable to access much needed loans which would help them grow and multiply output and profits.*

*New-age financial service providers in India are coming to terms with these needs of SMEs and have started rolling out innovative credit products with a focus on credit scoring methods which would be more suitable to the functioning of an SME than any general business. These default prediction models often entail the use of machine learning techniques like support vector machines or ensemble learning methods like random forest classifications, as seen in this study. The selection of independent variables which would contribute towards the model is another crucial aspect of the process. The focus has shifted from size of the business to the likelihood of the business entity to not default on future payments. The most important inferences from this study include insights about how the use of new credit default risk models would help accurately assess credit risk. Use of such techniques by credit lending institutions could effectively provide access to credit to a largely underserved market in India, and this could help SMEs and new businesses achieve growth and ride the manufacturing boom.*

Contents

[Abstract 6](#_Toc482762702)

[1 Introduction 9](#_Toc482762703)

[2 Literature Review 15](#_Toc482762704)

[3 Methodology and Analysis 19](#_Toc482762705)

[3.1 Model and Data 19](#_Toc482762706)

[3.2 Variables 20](#_Toc482762707)

[3.2.1 Loan Required 20](#_Toc482762708)

[3.2.2 Average Annual Turnover 20](#_Toc482762709)

[3.2.3 Length of Incorporation 20](#_Toc482762710)

[3.2.4 Present Year Profit 20](#_Toc482762711)

[3.2.5 Previous Year Profit 20](#_Toc482762712)

[3.2.6 Presence of Overdraft or Cash Credit Account 21](#_Toc482762713)

[3.2.7 Presence of Outstanding Loans 21](#_Toc482762714)

[3.2.8 Current Monthly EMI Payments 22](#_Toc482762715)

[3.2.9 Current Loans 22](#_Toc482762716)

[3.2.10 CIBIL Score 22](#_Toc482762717)

[3.2.11 Total Number of Accounts 23](#_Toc482762718)

[3.2.12 Days Past Due 23](#_Toc482762719)

[3.2.13 Liabilities/Total Assets 23](#_Toc482762720)

[3.2.14 Cash/Total Assets 24](#_Toc482762721)

[3.2.15 Return on Total Assets 24](#_Toc482762722)

[3.2.16 EBITDA/Interest Expenses 24](#_Toc482762723)

[3.3 Statistical Analysis 25](#_Toc482762724)

[3.3.1 Random Forest Classification 25](#_Toc482762725)

[3.3.2 Support Vector Machines 28](#_Toc482762726)

[4 Inferences and Results 33](#_Toc482762727)

[5 Conclusion 34](#_Toc482762728)

[6 References 35](#_Toc482762729)

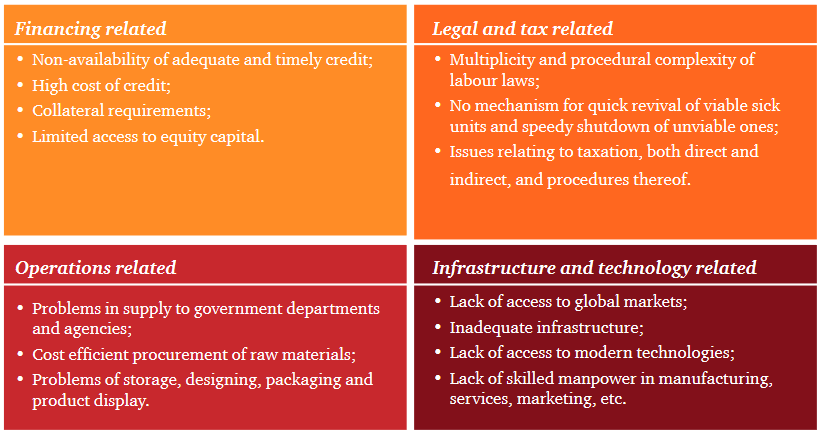
# 1 Introduction

The First Schedule to the Industries (Development and Regulation) Act, 1951 defines SMEs as:

* “In the case of the enterprises engaged in the manufacture or production of goods pertaining to any industry specified in, as –
  + a small enterprise is where the investment in plant and machinery is more than twenty-five lakh rupees but does not exceed five crore rupees
  + a medium enterprise is where the investment in plant and machinery is more than five crore rupees but does not exceed ten crore rupees”
* “In the case of the enterprises engaged in providing or rendering of services –
  + a small enterprise is where the investment in equipment is more than ten lakh rupees but does not exceed two crore rupees
  + a medium enterprise is where the investment in equipment is more than two crore rupees but does not exceed five crore rupees”

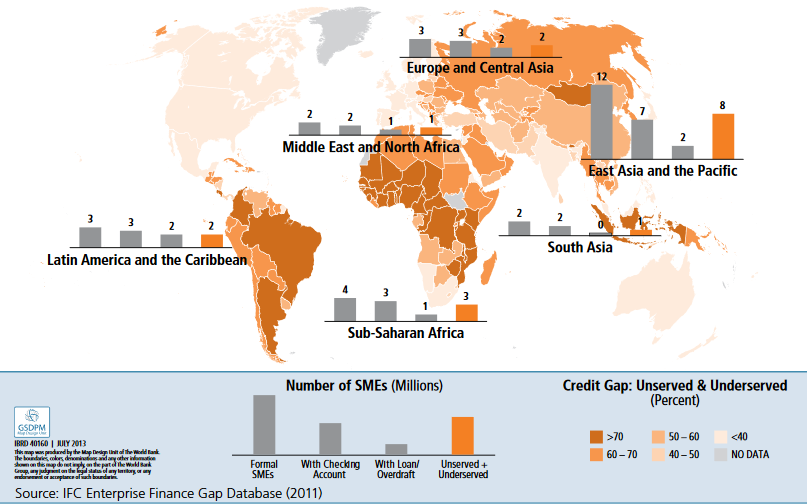
In India, companies that fall into the above mentioned categories are called SMEs or Small and Medium Enterprises. The Indian SME sector has emerged as one of the most dynamic sectors in the Indian economy over the last few decades. In 2011, SMEs contributed around 17% of India’s GDP. This figure is expected to grow to 22% in the year 2020. Around 40% of India’s workforce is employed in the SME sector and around 45% of India’s manufacturing output is credited to the SME sector. The sector by itself creates over 1 million jobs every year. SMEs are engines of economic and overall growth in India and they can uplift the poorest of the poor provided they are given ample support. Sustained and healthy growth of this sector is essential, since it is difficult to imagine rising overall living standards and social peace without such development.

The major challenges faced by SMEs in India are listed below:

Figure 1.1 Challenges faced by SMEs in India

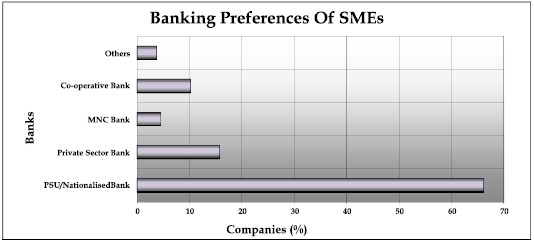
Source: PWC – Financing for MSMEs

Worldwide, there exists a large credit gap in SME financing.

Figure 1.2 Credit Gap (SMEs)

Source: IFC Enterprise Finance Gap Database

According to a Dun & Bradstreet Report, “The most preferred source of funding among companies in the sample was the public sector banks. Nearly 66% of the companies preferred banking with public sector banks (PSBs) or nationalized banks for their fund requirements, followed by private banks and cooperative banks. MNC banks were funding only 1% of the companies. There were approximately 4% companies that have funded their business through internal resources. In terms of availability of funds, some 7% of companies in the sample have claimed difficulty in acquiring funds. These are primarily North and West based, standalone companies, with a turnover of less than Rs 300 million.”

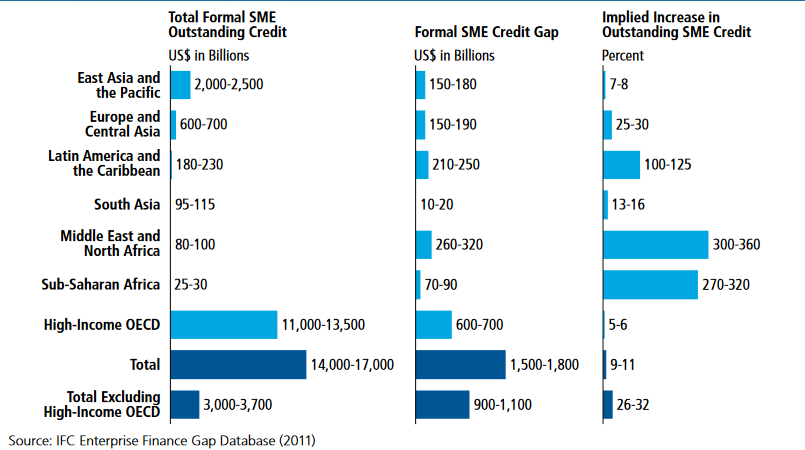
Figure 1.3 Banking Preferences of SMEs

Source: savings-banks.com

There exists a gap of approximately USD 126 billion in SME funding in India. This can be further classified into a debt gap of USD 84 billion and an equity gap of USD 42 billion.

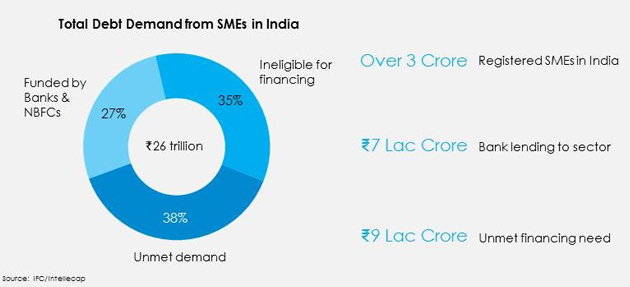
There is an asymmetry of information which exists in the SMEs in India. A large proportion of businesses are family owned. The amount of easily available information regarding sources and types of financing options are surprisingly limited. Funds availed from family and friends are the most popular sources of finance for SMEs in India. Money lenders both authorized and otherwise also form a major chunk of finance sources for these small businesses. These usually entail high interest rates and extremely uncertain terms.

Banks have long been attempting to close this gap in funding and provide financial support to SMEs in India. However, given the rigid procedures followed by banks, they have a long list of requirements which SMEs are more often than not able to fulfill. This includes extensive forms, documentation, collateral assessment etc. Banks try to mitigate risk to the greatest extent possible and end up cutting off financial support to the vast majority of SMEs.

Figure 1.4 Formal SME Credit Gap

Source: IFC Enterprise Finance Gap Database

In India:

Figure 1.5 Total Debt Demand from SMEs in India

Source: IFC Enterprise Finance Gap Database

To fill this unmet demand for credit, a variety of new loan products and innovative financing options are being introduced globally. In India, a new approach to SME financing is evolving wherein traditional methods of underwriting are being scrapped and firms are developing their own proprietary credit scoring mechanisms. The need for excessive paperwork and collateral is being reduced. This study attempts to critically look at a few innovative methods of credit risk scoring and how these would be useful in countering the current situation in SME credit lending in India.

# 2 Literature Review

Multiple studies have been conducted on how innovative statistical models and methods can be used to gauge credit risk. While a large number of these studies have been conducted on datasets from across the world, a considerable number of country specific studies too exist. The studies delve into how alternate methods of credit scoring and default prediction can help streamline credit lending processes thereby helping businesses and economies grow.

A large part of the efforts towards researching new methods of credit scoring has been on how to tailor these methods to suit businesses which have come up recently and as such lack the traditionally required factors to be deemed eligible for funding from credit lending entities.

Traditionally, bank loans, overdrafts, credit lines and credit cards have been the most common course of external finance for many SMEs (OECD Report on SME Financing, 2015). The borrower must pay a specified amount on interest to the lenders at regular intervals. This payment must be done regardless of how the company has been doing. The interest rate may or may not be changed with regard to the economy. The biggest problem that banks face while lending to SMEs is the asymmetry of information. SMEs do not often have proper financial statements which can be vetted. Their documents are often not credible enough to be considered legitimate proof for any lending activity to take place. Also, the distinction between the owner’s finances and those of the business is not clearly defined leading to confusion. Standards of corporate governance are usually not maintained and shareholders’ roles are not clearly defined. “At the one end of the risk/return spectrum are financing instruments that sustain the short and medium-to-long term financing needs of SMEs, but that rely on different mechanisms than traditional debt. This is the case of *asset-based finance*, such as *asset-based lending*, *factoring* and *leasing*, whereby a firm obtains cash, based not on its own credit standing, but on the value that a particular asset generates in the course of its business.”

* *“Trade credit* is also an important source of finance for many SMEs and start-ups, which can substitute or supplement short-term bank lending. This mainly consists of the extension of traditional credit instruments and credit-mitigation tools, such as loans and guarantees, to sustain import and export activities.”
* *“Corporate bonds*, *securitised debt* and *covered bonds*, in which investors in the capital markets, rather than banks, provide the financing for SMEs. While corporate bonds are direct instruments of debt finance for SMEs, securitisation and covered bonds represent indirect tools for supporting SME debt financing, in that the product issued to the firm is a loan.” (OECD Report on SME Financing, 2015)

“At the other end of the risk/return spectrum are financing instruments that enable an investor to accept more risk in exchange for a higher return, and are expected to produce a better alignment of the interests of certain kinds of SMEs and the providers of finance. *Hybrid instruments*, such as mezzanine finance, form a bridge between traditional straight debt and pure equity. *Seed and early stage finance* addresses the high risk-return segment of the business financing spectrum, boosting firm creation and development, whereas other equity-related instruments, such as *private equity* and *specialized platforms for SME public listing*, can provide financial resources for growth-oriented SMEs.” (OECD Report on SME Financing, 2015)

Leasing plays an important role in SME financing. SMEs do not raise money from capital markets and are usually dependent on bank financing. Banks’ appetite for risk is not very high and they also have other constraints. Leasing provides an alternate for SMEs to access credit. New enterprises which do not have collateral and/or a proven credit track record are provided access to credit in this manner. Securitization also helps to improve SMEs’ access to finance as it has a multiplier effect. “In addition to complementing private sector money, the participation of an institution like the EIF in these transactions can ‘crowd-in’ private resources via its positive signaling effect. An efficient public support will not only help to mitigate the impacts of the crisis but also to counteract market imperfections and weaknesses in SMEs’ access to finance which exist not only in times of crisis but on an on-going basis as a fundamental structural issue.” (Kraemer-Eis and Lang, 2015)

Finance for Agro-SMEs faces certain special problems as farmers are risky borrowers, small farmers more so (Global Partnership for Financial Inclusion, 2015). They have seasonal returns and cash flows aren’t very regular. Another issue is that their entire portfolio is subject to similar risks and that can create dire situations for them at times. “Agricultural finance in general involves higher transaction costs due to the greater distances, lower population densities and lower quality infrastructure encountered in serving rural areas, which discourages smaller transactions. Sub-optimal policy and regulatory environments can confuse the picture of value chain structure, pricing, profitability, and overall competitiveness. As with women entrepreneurs’ finance, limitations in laws and regulations governing movable assets financing raise the costs and risks of some possible solutions.” (Global Partnership for Financial Inclusion, 2015)

Since the credit crisis in Europe, demand for SME credit has shifted to short term working capital finance. Earlier demand was prominently for long term loans (Bain & Company, 2013). This has pushed for new methods of credit scoring for SMEs in Europe.

Only in Canada was the post 2010 access to credit for SMEs stable. It declined in almost all OECD nations (OECD Scoreboard, 2013).

In most developing countries, a dearth of credit for businesses is cited as one of the biggest obstacles to growth. This is the second most cited obstacle right after lack of uninterrupted electricity. Marketplace lending solutions offer hope to small businesses and investors alike (Global Agenda Council on the Future of Financing & Capital, 2015). Invoice finance platforms can be used effectively to overcome shortage of liquidity in the market. Existing supply chains can be integrated using supply chain finance in order to streamline processes.

The percentage of NPAs on loans taken by small firms is larger than that on loans taken by larger firms. Loan fees and interest rates are found to be higher in developing countries than in developed countries. The share of secured loans is higher among foreign owned banks than among domestically owned banks. Private banks are found to have a lower share of secured loans, greater fees and higher interest rates, when compared to government owned banks. (World Bank Development Research Group, 2013)

Banks are likely to enjoy 30% greater accuracy in default prediction if they use separate credit models for SMEs and corporates. Five financial ratios are found to be the most accurate predictors of SME default and can be used to develop a credit risk model specifically for SMEs. (Altman and Sabato, 2007)

Over a longer period of time, losses accumulate for SMEs and necessitate an increase in capital requirements. Internal credit risk models have a two-fold benefit. The first being that they take into account the correlations which might exist in retail portfolios and second being that they often lead to better capital allocation and improved loan pricing. (Dietsch and Petey, 2002)

In order to determine loan recovery rates, the most important drivers are loan size, collateral value, and year of loan origination (Chalupka and Kopecsni, 2008). Log-log models tend to perform better because of the asymmetric nature of the dependent variable.

Multivariate Discriminant Analysis (MDA) and Logistic Regression Models have been helpful in predicting bankruptcy of SMEs in Thailand over a period of 10 years, from 2001 to 2010. The logit model showed greater accuracy. A combined model which used both logit and MDA showed even higher accuracy. (Sirirattanaphonkun and Pattarathammas, 2012)

The most common predictors of default are financial ratios such as current ratio, leverage ratio, loans to total assets ratio, liquidity to profitability ratio etc. The higher the time duration of borrowing the lower is the likelihood of default. The likelihood of default deceases as the business entity’s size increase. However, it increases as the share of loans of a business’ assets increases. (McCann and McIndoe-Calder, 2012)

# 3 Methodology and Analysis

## 3.1 Model and Data

Based on the literature reviewed, the dependent variable was chosen to be a binary variable indicating acceptance or rejection for lending. A host of independent variables can be used in this case. These variables are chosen from among a number of credit rating bureau and business health variables which collectively act as a proxy for how well an SMEs business is functioning and how likely it is to not default on a loan.

The model used is as follows:

Y = I + C1X1 + C2X2 +……………………….……+ CnXn + e

Here,

**Y = Dependent Variable**

**I = Intercept**

**Cn = nth Constant Term (Coefficients of the Explanatory Variables)**

**Xn = nth Independent Variable**

**e = Error Term**

In this case, Y is the binary variable indicating good credit or bad credit and X1 through Xn are the credit rating and business health variables. The constant and error terms are results of the statistical modeling process and help to strengthen the model.

## 3.2 Variables

### 3.2.1 Loan Required

“The loan amount requested for by the applicant is an important factor for an SME credit default model.”

### 3.2.2 Average Annual Turnover

“Annual turnover is the percentage rate at which a mutual fund or exchange-traded fund replaces its investment holdings on an annual basis. Turnover is meant to adjust for the inflows and outflows of cash and report on the level of trading activity in the fund.

This value is calculated at least over the last two years. The time period for calculation may increase to the duration of the business’ operation since inception.”

### 3.2.3 Length of Incorporation

“The time duration between the current date and the date when the business entity was registered with a relevant and verified government registry is defined as the length of incorporation.”

### 3.2.4 Present Year Profit

“Total profits garnered by the business in the present year.”

### 3.2.5 Previous Year Profit

“Total profits garnered by the business in the previous year.”

### 3.2.6 Presence of Overdraft or Cash Credit Account

“An overdraft account is a loan arrangement under which a bank extends credit up to a maximum amount (called overdraft limit) against which a current (checking) account customer can write checks or make withdrawals. The most common form of business borrowing, an overdraft is a type of revolving loan where deposits (credits) are available for re-borrowing, and interest is charged only on the daily overdraft (debit) balance. It is, however, also a demand loan: the facility can be cancelled (and entire outstanding amount 'called') at any time by the lender at its discretion, without any warning notice or explanation.”

“A line of credit is credit source extended to a government, business or individual by a bank or other financial institution. A line of credit may take several forms, such as overdraft protection, demand loan, special purpose, export packing credit, term loan, discounting, purchase of commercial bills, traditional revolving credit card account, etc. It is effectively a source of funds that can readily be tapped at the borrower's discretion. Interest is paid only on money actually withdrawn. (However, the borrower may be required to pay an unused line fee, often an annualized percentage fee on the money not withdrawn.) Lines of credit can be secured by collateral, or may be unsecured.”

Presence or absence of an overdraft account and/or a cash credit line is a binary variable.

### 3.2.7 Presence of Outstanding Loans

The presence of any outstanding loans under the accounts held by the business entity is a numeric variable.

### 3.2.8 Current Monthly EMI Payments

“An EMI is a fixed payment amount made by a borrower to a lender at a specified date each calendar month. Equated monthly installments are used to pay off both interest and principal each month, so that over a specified number of years, the loan is paid off in full.”

### 3.2.9 Current Loans

All current loan amounts being held by the applicant add up to this.

### 3.2.10 CIBIL Score

“The Credit Information Bureau (India) Ltd, better known as CIBIL, is the premier agency for providing credit reports and scores pertaining to individuals. CIBIL sources financial data of individuals such as loan and credit card information from leading banks and other financial institutions in India. This data is then presented in the form of a CIBIL credit report, also known as a Credit Information Report (CIR). CIBIL was incorporated in 2000 and has continued to expand its presence throughout the country. It is backed by TransUnion International and Dun and Bradstreet, which are major global credit bureaus and agencies.”

“CIBIL credit score or credit rating is, in short, a 3-digit figure ranging from 300 to 900 points which denotes a person’s credit worthiness. This figure is derived from the CIBIL report of an individual through advanced statistical algorithms that take into account their credit history including borrowings, repayment patterns, defaults in repayment and other data relevant to a person’s creditworthiness. CIBIL scores are calculated on the basis of at least 6 months of historical financial data of an individual. The data is fed into an algorithm with 258 different variables; each with a different weightage.”

### 3.2.11 Total Number of Accounts

The total number of bank accounts, both credit card and non-credit card, held by the business entity and its promoters.

### 3.2.12 Days Past Due

“Past due is a loan payment that has not been made as of its due date. A borrower who is past due may be subject to late fees, unless the borrower is still within a grace period. Failure to repay a loan on time could have negative implications for the borrower's credit status or cause the loan terms to be permanently adjusted.”

“Days Past due or DPD means, that for any given month, how many months’ worth of payment is unpaid. And this information is for each account. For each account DPD information can be seen for each month for the last 3 years i.e. 36 months.”

“A CIBIL report contains the past 36 months of credit information. Each and every month, a lender who is a member of CIBIL, will update the CIBIL with latest information like timely payments made, outstanding loans etc. So each month, a new month’s data is added and the oldest month (36th month) is removed from the CIBIL report and this way a sliding window of 36 months’ data is available on your CIBIL report at any given point of time.”

### 3.2.13 Liabilities/Total Assets

“Debt Ratio is a financial ratio that indicates the percentage of a company's assets that are provided via debt. It is the ratio of total debt (the sum of current liabilities and long-term liabilities) and total assets (the sum of current assets, fixed assets, and other assets such as 'goodwill').”

### 3.2.14 Cash/Total Assets

“The cash asset ratio is the current value of marketable securities and cash, divided by the company's current liabilities. Also known as the cash ratio, the cash asset ratio compares the dollar amount of highly liquid assets (such as cash and marketable securities) for every one dollar of short-term liabilities. This figure is used to measure a firm's liquidity or its ability to pay its short-term obligations. Ideal ratios will be different for different industries and for different sizes of corporations, and for many other reasons.”

### 3.2.15 Return on Total Assets

“Return on total assets (ROTA) is a ratio that measures a company's earnings before interest and taxes (EBIT) against its total net assets. The ratio is considered an indicator of how effectively a company is using its assets to generate earnings before contractual obligations must be paid.”

### 3.2.16 EBITDA/Interest Expenses

“The EBITDA-to-interest coverage ratio is a ratio that is used to assess a company's financial durability by examining whether it is at least profitably enough to pay off its interest expenses. A ratio greater than 1 indicates that the company has more than enough interest coverage to pay off its interest expenses.”

## 3.3 Statistical Analysis

### 3.3.1 Random Forest Classification

A random forest is an ensemble learning method which employs multiple decision trees over subsets of the original dataset to arrive at classifications which are then averaged out over the entire population. The output is in the form of a classification or a mode.

A number forest of decision trees, is built on bootstrapped training samples. But when building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors. The number of predictors considered at each split is approximately equal to the square root of the total number of predictors. In other words, in building a random forest, at each split in the tree, the algorithm is not even allowed to consider a majority of the available predictors.

Random forests make sure that each predictor considers only a subset of the total set. This adjusts for any correlations which might exist within among the predictors. Averaging out multiple predictors which are correlated to each other might substantially skew results.

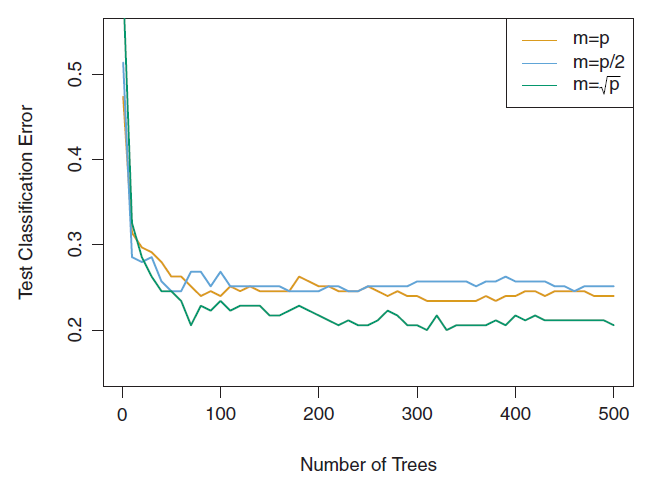


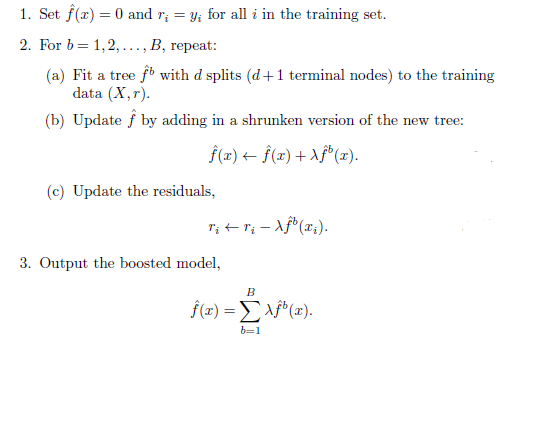
Figure 3.1 Test Classification Error in Random Forest Classification

Source: Springer Texts in Statistics

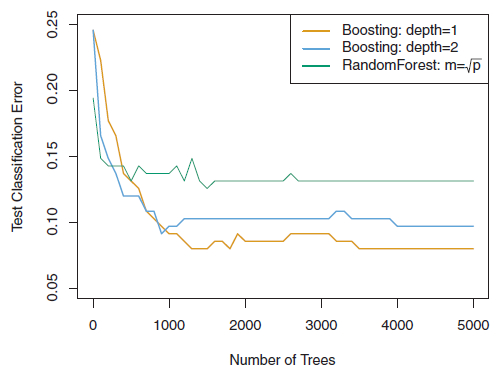
As seen in the image above, the selection of subsets using a square root of the total number of predictors allows for a lower value of classification error than we would otherwise have.

The prediction results from random forest decision trees can be improved by using a method called boosting. Instead of fitting a single large decision tree to the data, which might cause overfitting, the boosting approach lets the model learn slowly. The decision tree is fitted to the residuals from the model. Instead of the outcome Y, the trees are fitted to the current residuals. This new decision tree is then added to the already fitted function and then the results are updated. Each of these trees is small and has only a few terminal nodes, instead of one large tree with multiple nodes. The number of nodes is determined by the parameter d in the algorithm.

The algorithm for boosting on random forests can be noted as follows:



The depth of boosting can be tweaked in order to further decrease the classification error compared to using a normal random forest model:

Figure 3.2 Random Forest Error vs Boosting Error

Source: Springer Texts in Statistics

### 3.3.2 Support Vector Machines

“Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression models. However, it is mostly used in classification problems. In this algorithm, each data item is plotted as a point in an n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate.”

The support vector machine is a generalization of a simple and intuitive classifier called the maximal margin classifier. This classifier however cannot be applied to common data sets as it requires proper separation of the classes by a linear boundary. The support vector classifier is an extension of the maximal margin classifier that can be applied to data sets which have non liner class boundaries. Support vector machines are used for binary classification in data sets which have two classes.

In a p-dimensional space, a hyperplane is a flat affine subspace of dimension p − 1. For instance, in two dimensions, a hyperplane is a flat one-dimensional subspace—in other words, a line. In three dimensions, a hyperplane is a flat two-dimensional subspace—that is, a plane. In p > 3 dimensions, it can be hard to visualize a hyperplane, but the notion of a (p − 1)-dimensional flat subspace still applies.

A hyperplane can be mathematically defined as:

β0 + β1X1 + β2X2 + . . . + βpXp = 0

An example of a two dimensional hyperplane which divides the data set into two halves:

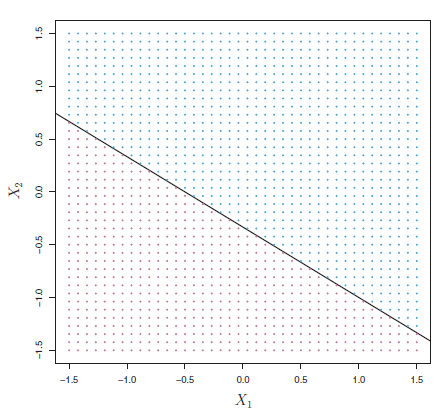


Figure 3.3 Hyperplane

Source: Springer Texts in Statistics

If a number of separating hyperplanes are created in order to divide a dataset, the one that is farthest from the training observations is called the maximal margin classifier. The smallest perpendicular distance from a training observation to the hyperplane is called the margin.

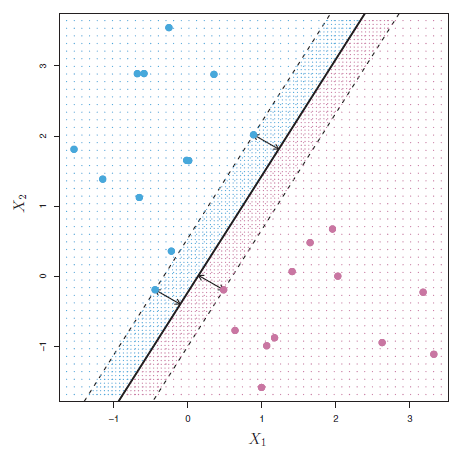


Figure 3.4 Maximal Margin Classifier

Source: Springer Texts in Statistics

Rather than seeking the largest possible margin so that every observation is not only on the correct side of the hyperplane but also on the correct side of the margin, the structured vector classifier instead allows some observations to be on the incorrect side of the margin, or even the incorrect side of the hyperplane.

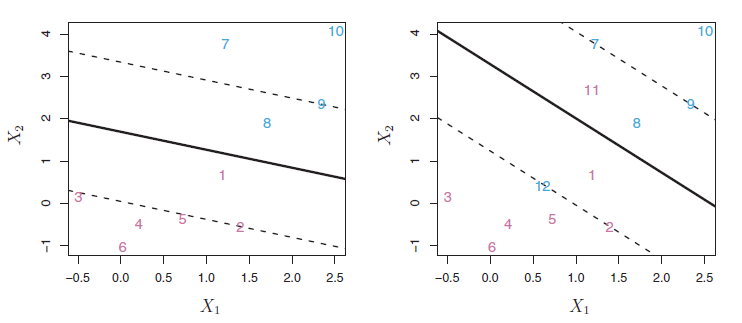
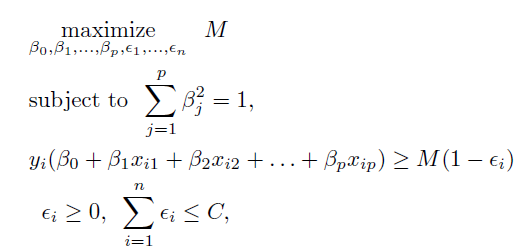


Figure 3.5 Support Vector Classifier

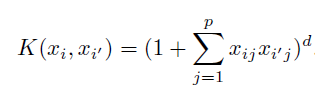
Source: Springer Texts in Statistics



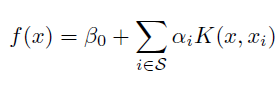
Here, C is the tuning parameter which controls the degree of tradeoff between bias and variance for the support vector classifier.

“The support vector machine (SVM) is an extension of the support vector machine classifier that results from enlarging the feature space in a specific way, using kernels. A kernel is a function that quantifies the similarity of two observations.”

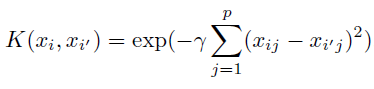
A linear kernel:



A non-linear kernel:



A polynomial kernel:



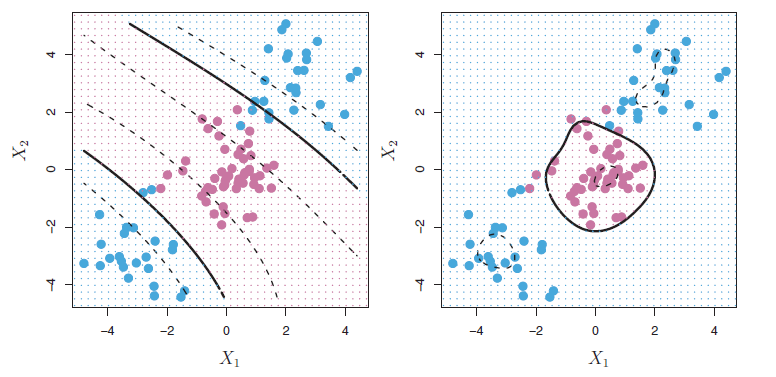


Figure 3.6 Support Vector Machines

Source: Springer Texts in Statistics

# 4 Inferences and Results

The major advantage that random forests have over SVMs is that they do not expect linear features or even features that interact linearly. A random forest model can easily handle categorical (binary) features. The other main advantage is that, because of how they are constructed (using bagging or boosting) random forest algorithms handle very well high dimensional spaces as well as large number of training examples.

In practice, an SVM with a linear kernel is not very different from a Logistic Regression. SVMs are particularly useful when the dataset is not linearly separable. In that case, an SVM with a non-linear kernel can be used. Another related reason to use SVMs is the prevalence of highly dimensional spaces which are more conducive to implementing an SVM. For example, SVMs have been reported to work better for text classification. Unfortunately, the major downside of SVMs is that they can be painfully inefficient to train. Hence, they can be hard to implement for industry level modeling problems.

Random forests are particularly adept when it comes to feature selection for the model. A support vector machine on the other hand relies on the input to be pre-molded and features to be selected prior to the modeling process. Feature selection is probably the trickiest part of the process and as such the random forest performing it on its own lends to its popularity among credit default modelers these days.

Thus, keeping the variables discussed in mind, a random forest decision tree would ideally be the best choice to construct a credit scoring or default prediction model for SMEs in India.

# 5 Conclusion

The finance industry in India is currently undergoing a sea change. Banks are making efforts to go digital in order to reach out to a newly online population. Individuals and businesses alike are realizing the power of credit in helping growth. Speaking about disruption in financial services, Mr. Nandan Nilekani referred to the current financial services industry in India as undergoing a ‘WhatsApp like movement’, one that would fundamentally change the way the industry works and how services are rendered. The untapped SME market which has been reeling from a dearth of adequate financial services is huge in India currently. This fuels the need for innovative methods of credit scoring and default risk estimation which would enable efficient handling of credit demand from the industry. Traditional credit risk estimation models do not work efficiently when it comes to SMEs because the fundamentals being looked at in both cases are very different. Traditionally, loans have been given out on the basis of collateral provided and general health of the business. The requirement for documentation is also extensive in most cases. Processing speeds are excruciatingly slow and interest rates and terms are often non-negotiable. On the other hand, new-age lenders tend to focus on the future and look at the likelihood of a business to be able to pay off installments in the future. Even though interest rates may be slightly higher, the ease of getting a loan from such lenders works well for SMEs. These innovative methods of credit risk scoring using machine learning techniques allows for accurate scoring and quick disbursal of money. This in turn propels SMEs forward and also allows the credit lending institutions to function efficiently. Thus, the disruption in India’s financial services sector depends on how well lenders are able to adapt to changing definitions of good or bad credit and how well they are able to assess credit risk using innovative techniques.

# 6 References

1. 'New Approaches to SME and Entrepreneurship Financing: Broadening the Range of Instruments', OECD. (2015)
2. Kraemer-Eis, H. and Lang, F. in 'The importance of leasing for SME finance', Working Paper 2012/15, European Investment Fund. (2012)
3. 'Small and Medium Enterprise Finance: New Findings, Trends and G-20/Global Partnership for Financial Inclusion Progress', Global Partnership for Financial Inclusion, International Finance Corporation. (2013)
4. 'Restoring Financing and Growth to Europe's SMEs', Bain and Company. (2013)
5. 'Financing SMEs and Entrepreneurs: An OECD Scorecard', OECD. (2013)
6. 'The Future of FinTech: A Paradigm Shift in Small Business Finance', Global Agenda Council on the Future of Financing & Capital, World Economic Forum. (2015)
7. Beck, T., Demirguc-Kunt, A. and Peria, M.S.M. in 'Bank Financing for SMEs around the World: Drivers, Obstacles, Business Models, and Lending Practices', Development Research Group, The World Bank. (2008)
8. Sabato, G. and Altman, E. in 'Modeling Credit Risk for SMEs: Evidence from the US Market', Journal 43/3, Blackwell Publishing Asia. (2007)
9. Petey, J. and Dietsch, M. in 'The credit risk in SME loans portfolios: Modeling issues, pricing, and capital requirements', Journal of Banking & Finance 26/303–322, Elsevier BV. (2002)
10. Chalupka, R. and Kopecsni, J. in 'Modelling Bank Loan LGD of Corporate and SME Segments', IES Working Paper: 27/2008. (2008)
11. Sirirattanaphonkun, W. and Pattarathammas, S. in 'Default Prediction for Small-Medium Enterprises in Emerging Market: Evidence from Thailand', Seoul Journal of Business Volume 18, Number 2. (2012)
12. McCann, F. and McIndoe-Calder, T. in 'Determinants of SME Loan Default: The Importance of Borrower-Level Heterogeneity', Research Technical Paper, Central Bank of Ireland. (2012)
13. Ghodselahi, A. in 'A Hybrid Support Vector Machine Ensemble Model for Credit Scoring', International Journal of Computer Applications (0975 – 8887). (2011)
14. Zhang, A. in 'Statistical Methods in Credit Risk Modeling', University of Michigan. (2011)'New Approaches to SME and Entrepreneurship Financing: Broadening the Range of Instruments', OECD. (2015)
15. <https://en.wikipedia.org>
16. <http://www.investopedia.com>
17. <https://www.researchgate.net>
18. <http://www.defaultrisk.com>
19. <https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm>
20. <https://www.bankbazaar.com/cibil/cibil-credit-score.html>