Contents

List of Tables iii

List of Figures iv

Nomenclature vi

Abstract vii

1 Introduction 1

1.1 Applications . . . 1
1.2 Motivation . . . 2
1.3 Objectives . . . 2
1.4 Organization of Report . . . 2

2 Theoretical Background 3

2.1 Machine Learning Algorithms . . . 3
2.1.1 Logistic Regression (LR) . . . 3
2.1.2 Decision Tree (DT) . . . 4
2.1.3 Random Forests (RF) . . . 4
2.1.4 Support Vector Machine (SVM) . . . 5
2.1.5 K-Nearest Neighbours (KNN) . . . 6
2.1.6 Bayesian Network (BN) . . . 6
2.1.7 Linear Discriminant Analysis (LDA) . . . 7
2.1.8 Artificial Neural Networks (ANN) . . . 8
2.2 Overview Banking and Risk Management Terminologies . . . 9
2.2.1 Risk . . . 9
2.2.2 Credit Risk . . . 9
2.2.3 Market Risk . . . 9
2.2.4 Liquidity Risk . . . 10
2.2.5 Operational Risk . . . 10
2.3 Summary . . . 10

3 Literature Review 11

3.1 Credit Risk 11
3.1.1 Traditional Machine Learning Algorithms for Credit Scoring 11
3.1.2 Neural Networks for Credit Scoring 11
3.2 Market Risk 12
3.3 Liquidity Risk 13
3.4 Operational Risk 13
3.5 Summary 16

4 Case Study in Credit Risk **17**
4.1 German Credit Dataset 17
4.2 Exploratory Data Analysis (EDA) 17
4.2.1 Distribution of Good and Bad Credit 17
4.2.2 Age Distribution 18
4.2.3 House Type Distribution 19
4.2.4 Loan Purpose Distribution 19
4.3 Results 20

5 Conclusion **22**

References **23**

Acknowledgment **25**

List of Tables

1 Literature Survey Summary 14
2 Performance of classifiers on German Credit Dataset 20

List of Figures

1 Logistic function curve 3
2 Hyperplane construction in SVM 5
3 Kernel mapping to linear boundaries 5
4 Example of KNN 6
5 Example of BN 7
6 Fisher's LDA Visualization 7
7 Single neuron in an ANN 8
8 Distribution of good and bad credit 18
9 Distribution of credit based on age 18
10 Distribution of credit based on purpose 19
11 Distribution of credit based on purpose 19
12 ROC for classifiers 21

Nomenclature

English Alphabets

e	Euler's constant	2.71828

Non-English Alphabets

α	Alpha
β	Beta
ϵ	Epsilon
ω	Omega
σ	Sigma
$\sum$	Uppercase Sigma

Abbreviations

ANN	Artificial Neural Network
$ARCH$	Autoregressive Conditional Heteroscedasticity
AUC	Area Under Curve
BN	Bayesian Network
$CART$	Classification and Regression Trees
DAG	Directed Acyclic Graph
DT	Decision Tree
EDA	Exploratory Data Analysis
$GARCH$	Generalized Autoregressive Conditional Heteroskedasticity
KNN	K-Nearest Neighbours
LCR	Liquidity Coverage Ratio

LDA Linear Discriminant Analysis

LR Logistic Regression

MLP Multilayer Perceptron

MSE Mean Squared Error

NPA Non Performing Assets

$NSFR$ Net Stable Funding Ratio

RBI The Reserve Bank of India

RF Random Forest

ROC Receiver Operating Characteristics

SVM Support Vector Machine

VAR Value at Risk

Abstract

Banking forms the backbone of modern society. With banks serving millions of customers from every corner of the world, it becomes crucial for banks to stay competitive, sustainable and at the same time work under compliance constraints. Risk management is an area of study which aims to reduce fraud, improve processes and decrease losses to the banking industry. Several state-of-the-art machine learning and statistical models are used to identify, analyze and mitigate risks for the organizations. These models are exceedingly improved by the terabytes of data that banks generate daily. This report presents a comprehensive analysis of various techniques used for risk management in the banking sector. A comparative study is performed on various machine learning models and their strengths for different risk management tasks. Finally, a case study is also presented highlighting the complete process of a sample risk management task for credit scoring.

Keywords: *Banking - Risk Management - Machine Learning - Credit Score*

1 Introduction

The banking industry is an essential part of the global economy handling over 150 trillion U.S Dollars worth of global assets. Amidst the economic recession caused due to the COVID-19 pandemic, the total value of banking-related fraud in India doubled from Rs 71,534 Crores in 2018-19 to Rs 1,38,422 Crores in the fiscal year 2020-21 [1]. Global recessions like the 2008 Financial Crisis and the recent COVID-19 pandemic slows down economic output and put a lot of pressure on financial institutions, especially banks. Having a strong infrastructure to manage and mitigate financial risk becomes crucial to keeping the global economic engine running.

Over the last few decades, advancements in Machine Learning have been extensively utilized to help in risk management tasks. With almost all financial instruments running in a digital form, machine learning models are employed extensively to detect fraud, predict commodity prices, track inflation etc. A lot of research has been done academically and in the industry since the 2008 Financial Crisis to build state of the art models for various risk management activities. This report presents a summary of multiple machine learning techniques that are used to solve a variety of risk management related problems.

1.1 Applications

Risk management is a 17.1 Billion U.S. Dollar industry responsible for handling inflation, financial frauds, cyber threats, market volatility etc. [2]. Machine Learning models significantly improve the accuracy of existing risk management tasks and can do so globally by working on petabytes of data that the banking sector produces daily. Typical applications of machine learning in risk management tasks include:

- **Credit Risk:** Determining if the borrower can repay the loan.
- **Market Risk:** Predicting volatility and movement in commodity and equity markets.
- **Financial Fraud Risk:** Identifying money laundering patterns from transactions.
- **Liquidity Risk:** Identifying stability of sectors before investing in financial instruments
- **Operational Risk:** Detecting Fraud and suspicious transactions

1.2 Motivation

Risk Management is essential for a country's economy. The Reserve Bank of India (RBI) reported the presence of about Rs. 900 Thousand Crores of Non-Performing Assets (NPA) in 2020 in the Indian banking sector [3]. These are essentially bank loans that are in default or arrears. This amount of loss is roughly equivalent to the cost to build an international airport in India. Machine Learning plays a monumental role in reducing exposure to such risks, which can significantly impact national and global economies.

1.3 Objectives

This report aims to summarize the work done in risk management in the banking sector from a machine learning point of view. Additionally, this report also presents a case study comparing the performance of different machine learning models to evaluate credit risk on a single dataset.

1.4 Organization of Report

This report is subdivided into five chapters. The first chapter gives an overview of the banking industry along with the applications, motivations and objectives of this report. The second chapter provides the necessary theoretical background on machine learning algorithms and concepts related to banking and risk management. The third chapter presents a comparative study of various academic work done for different risk management tasks using machine learning. The fourth chapter is a case study for a particular risk management problem that builds upon the previous chapters and reports the results of various machine learning algorithms. The final chapter concludes all the findings of the seminar report in a concise format.

2 Theoretical Background

This chapter is subdivided into two parts. The first part is a brief outline of standard machine learning models and terminologies. The second part explains various terminologies related to finance, banking and risk management.

2.1 Machine Learning Algorithms

2.1.1 Logistic Regression (LR)

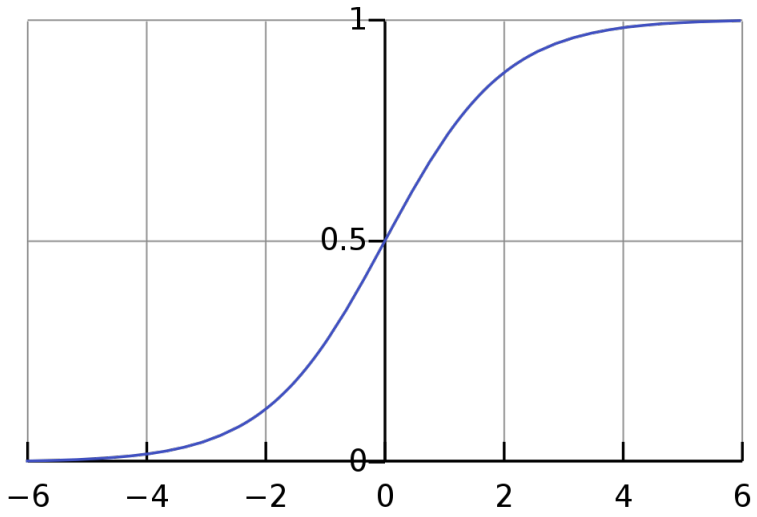

Fig. 1. Logistic function curve

In statistics, LR models the probability of a specific class or event happening. This probability estimate is further used for binary classification and with minor extensions to perform multi-class classification. The algorithm is based on a logistic function which is a sigmoidal curve. It takes a real-valued input and outputs a number between zero and unity (Fig. 1). This output is interpreted as the probability of the task. Mathematically, the logistic function is represented as follows:

$$\sigma(t) = \frac{e^t}{1 + e^t} = \frac{1}{1 + e^{-t}} \tag{1}$$

If t is derived from a single independent linearly varying variable x, the generalized logistic function can be represented as follows:

$$t = \beta_0 + \beta_1 x \tag{2}$$

$$p(x) = \sigma(t) = \frac{1}{1 + e^{\beta_0 + \beta_1 x}} \tag{3}$$

2.1.2 Decision Tree (DT)

Decision Tree Learning involves the construction of DTs, which will be traversed to conclude a given set of observations of data points. Classification Trees are models where the predicted outcome is discrete. Decision Tree training algorithms work in a top-down fashion to split the data set at every stage using a specific metric. Popular decision trees and their splitting metrics are discussed below.

Classification and Regression Tree (CART)

CART model uses the Gini Impurity Index I_G as its splitting parameter, which is computed for j classes as follows:

$$I_G(p) = 1 - \sum_{i=1}^{j} p_i^2 \tag{4}$$

Information Gain based DTs

DT models like ID3, C4.5 and C5.0 use information gain, which is based on the entropy of the dataset. Mathematically Entropy H and Information Gain IG for j classes are defined as below:

$$H(T) = -\sum_{i=1}^{j} p_i \log_2 p_i \tag{5}$$

$$IG(T, a) = H(T) - H(T|a) \tag{6}$$

2.1.3 Random Forests (RF)

RFs are constructed using ensemble learning on DTs. The majority output of the decision trees is considered the output of an RF in a classification task. The mean of all outputs is taken for a regression task. RFs are used as a corrective measure for overfitting of DT based models.

2.1.4 Support Vector Machine (SVM)

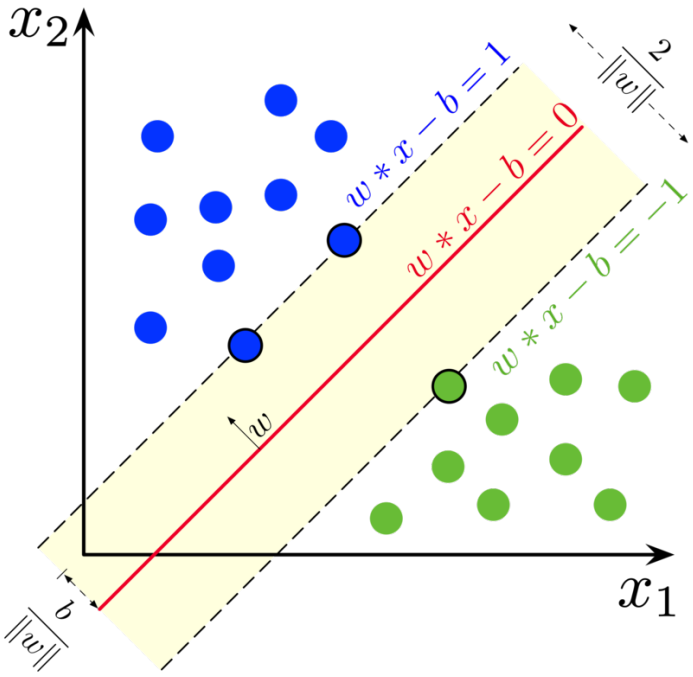

Fig. 2. Hyperplane construction in SVM

SVMs are supervised learning models used for regression and classification tasks. For classification, two hyperplanes are constructed, separating linearly separable data classes. The hyperplane at the midway of these two planes is the maximum hyperplane which acts as a classification boundary (Fig. 2).

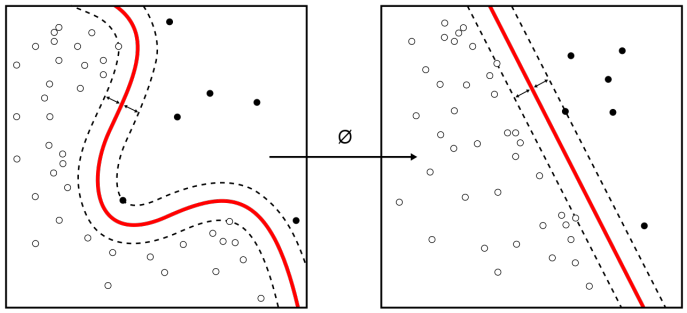

Fig. 3. Kernel mapping to linear boundaries

For non-linear data, a kernel is used, which is a function mapping the data into higher dimensional linear feature space (Fig. 3)

2.1.5 K-Nearest Neighbours (KNN)

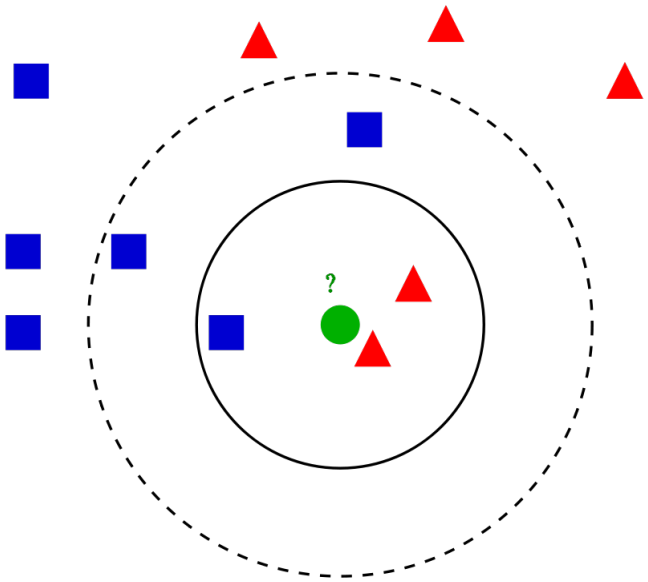

Fig. 4. Example of KNN

KNN is a supervised algorithm for supervised classification and regression tasks. KNN works by finding the K closest data points to a given query point. The output is the most frequent label for classification and the mean of the label for regression tasks. The value of K is user-defined as a hyperparameter, and the distance metric is euclidean or hamming distance depending on the context of the dataset.

2.1.6 Bayesian Network (BN)

A BN is a Directed Acyclic Graph (DAG) used to represent the probabilistic dependencies of a collection of random variables. Each node of the BN represents a random variable, and each edge between the nodes represents the conditional probability between the nodes, which is derived from the Bayes theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{7}$$

As an example, consider the BN in (Fig. 5), modelling the conditions when the grass can be wet. If it rains or the sprinklers are active, there is a higher probability of grass being wet. Moreover, if it is raining, the probability of the sprinkler being kept active is lesser.

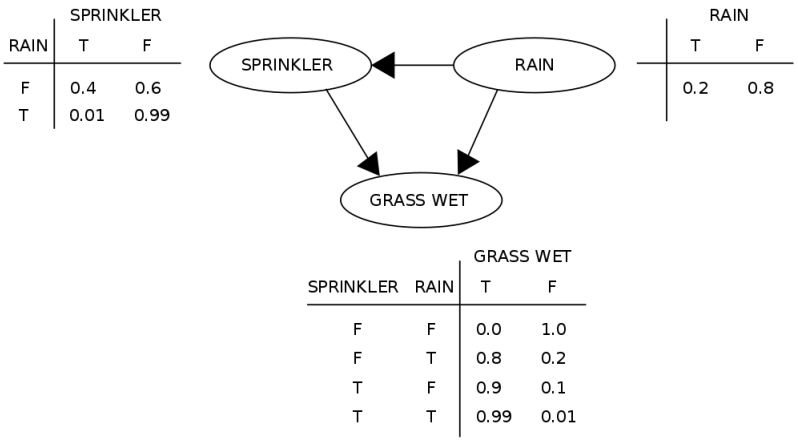

Fig. 5. Example of BN

2.1.7 Linear Discriminant Analysis (LDA)

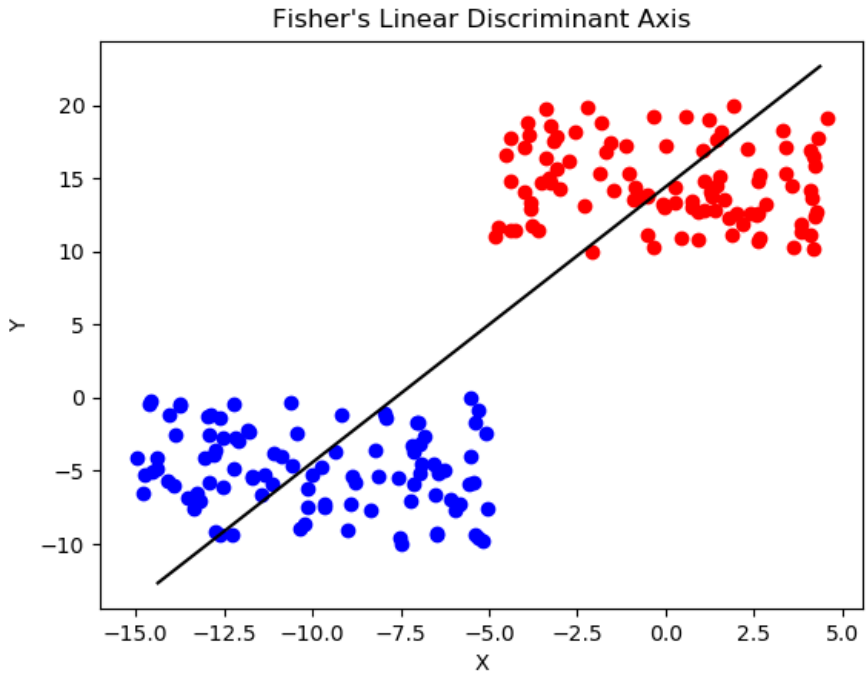

Fig. 6. Fisher's LDA Visualization

LDA is used for dimensionality reduction, especially for supervised classification tasks to separate two or more classes spatially. LDA takes a component of the data points on a single axis tuned to maximize the distance between two classes and minimize the variation within members of a single class.

2.1.8 Artificial Neural Networks (ANN)

ANNs are machine learning models mimicking biological neural networks. An ANN consists of neuron layers where each layer contains an array of artificial neurons. These neurons are essentially mathematical functions that take input values from previous layers and are tunable using a set of weights and biases. ANNs are used for supervised, unsupervised as well as reinforcement learning tasks.

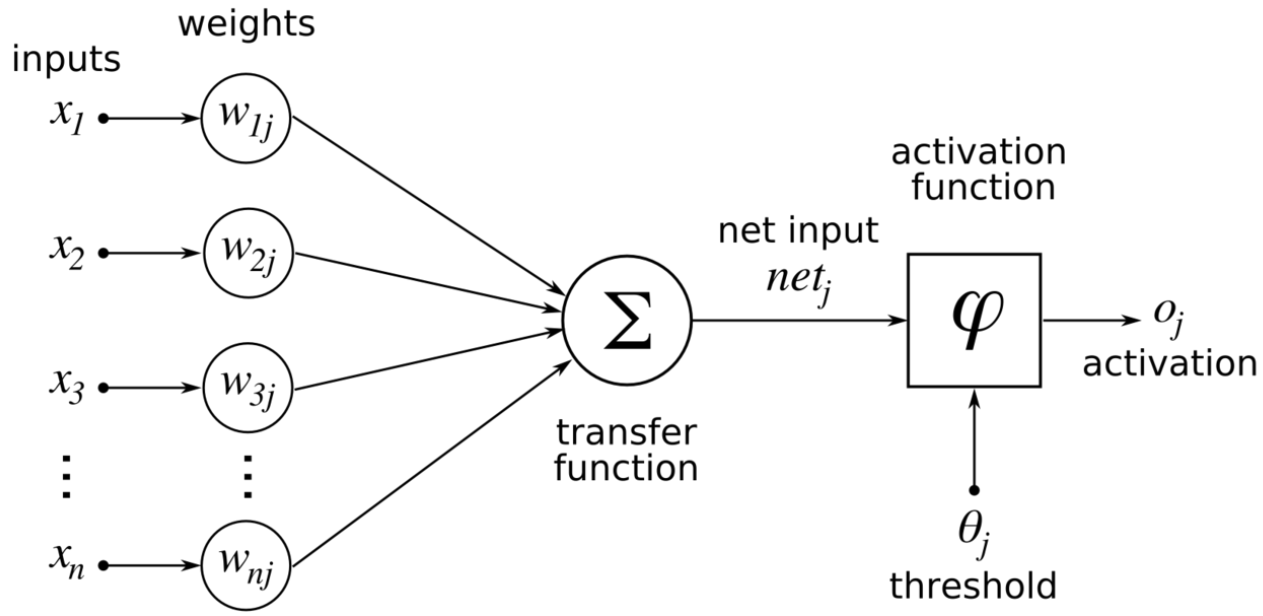

Fig. 7. Single neuron in an ANN

Multilayer Perceptron (MLP)

MLPs are a class of feedforward ANNs where the connections between layers of neurons are unidirectional and acyclic. Each neuron is an activation function like hyperbolic tan, logistic function and rectifier function.

$$y(v_i) = \tanh(v_i) \tag{8}$$

$$y(v_i) = \frac{1}{1 + e^{v_i}} \tag{9}$$

$$y(v_i) = \max(0, v_i) \tag{10}$$

2.2 Overview Banking and Risk Management Terminologies

2.2.1 Risk

Risk in finance is defined as uncertainty or volatility of unexpected outcomes, representing the value of assets, equity, or earnings. This includes both positive as well as negative outcomes. A higher return is associated with a more significant variability of outcomes [4]. Risk is broadly classified into Credit Risk, Market Risk, Liquidity Risk and Operational Risk.

2.2.2 Credit Risk

Credit risk is the risk of defaulting on bank loans or failing to comply with the debt obligations of a borrower [5]. Credit risk also refers to the decline of a borrower's credit quality, which does not directly indicate default but involves a higher probability of default. The book value does not change when the borrower's credit quality diminishes, but its economic value is lower because the probability of default increases [6].

2.2.3 Market Risk

Market risk is the risk of the movement of market prices. Prices of market parameters that fluctuate randomly include equity indexes, commodity prices, interest rates and foreign exchange rates [6].

Time Series Forecasting

Time Series Forecasting is used to predict the market price movement. It involves conceptualizing statistical and machine learning models to predict future values based on previously observed values.

Innovation

Innovation is the difference between a forecast made by a statistical algorithm and the actual value for a time-varying function.

Heteroscedasticity

Heteroscedasticity is the variance or volatility observed in a time-varying random function.

2.2.4 Liquidity Risk

Liquidity is the ease with which a financial instrument can be exchanged for money without losing value. High liquidity indicates a high supply and demand for an asset, which means there will always be buyers and sellers. Liquidity risk is the uncertainty linked with the outcome and possible losses of such an asset transaction into funds [7].

Measuring Liquidity

Liquidity is measured using ratios like Liquidity Coverage Ratio (LCR) and Net Stable Funding Ratio (NSFR). LCR measures the number of liquid assets a bank has to cover for any short term liquid fund requirements. NSFR, on the other hand, measure a bank's long term resilience to liquidity risks. Basel Norms require that both the ratios must be maintained above 100% by banks.

$$LCR = \frac{\text{stock of high quality liquid assets}}{\text{total net cash} - \text{flow over the next 30 days}} \tag{11}$$

$$NSFR = \frac{\text{available amount of stable funding}}{\text{required amount of stable funding}} \tag{12}$$

2.2.5 Operational Risk

Operational Risk is the risk associated with failures in internal procedures, external events, and infrastructure. Fraud detection, cyber security attacks, suspicious transaction detection, money laundering all come under the radar of operational risk [6].

Money Laundering

Money Laundering converts substantial amounts of funds earned from crimes and terrorism into origination from a legitimate source. This is accomplished by routing transactions through numerous layers of legal transactions to hide the natural source of income.

2.3 Summary

In conclusion, this chapter briefly introduces various concepts and terminologies that would be a prerequisite for Chapter 3 and Chapter 4. Machine learning algorithms are used to solve a vast array of economics, banking, and risk management problems.

3 Literature Review

This chapter discusses various contributions of researchers and scholars for Risk Management in Banking using Machine Learning Models.

3.1 Credit Risk

Credit Risk analysis is the fundamental problem of the lending economy, with initial efforts dating back to the 20th century. The financial recessions of the 21st century and the ever-increasing complexity of regulatory constraints have led to increased academic and business efforts towards the problem. From a Machine Learning perspective, evaluating credit risk involves binary classification: Determining if a potential customer is credit-worthy from their financial history.

3.1.1 Traditional Machine Learning Algorithms for Credit Scoring

A family of statistical frameworks have been adopted since the early 1900s for credit scoring potential borrowers. Among the traditional classifiers, SVMs have yielded significantly better results.

Bellotti et al. tested the performance of SVMs for credit score classification and contrasted the results with traditional approaches [8]. The paper compares the performance of different algorithms like LR, SVM, LDA and KNN. The work concluded with LDA and SVM reporting the highest AUC scores for the ROC curves when trained on a dataset of about 25,000 over three months of 2004 for credit card based lending [8]. However, SVMs are very sensitive to the outliers of the datasets. Wang et al. introduced a fuzzy SVM algorithm for the classification task to reduce this sensitivity and increase the generalization of the data [9].

3.1.2 Neural Networks for Credit Scoring

Early works in the area primarily included the use of SVM based algorithms. However, with larger datasets, training becomes computationally expensive [10]. ANNs have been widely researched for credit scoring [11, 12]. Several Ensemble methods have been employed to improve the accuracy of the ANN models [13].

Angelini et al. used ANNs to determine credit risk among small businesses as credit borrowers [14]. The work proposes a real-world dataset of Italian businesses and provides a comprehensive network performance analysis for various neuron configurations.

Tsai et al. employed a single MLP classifier as a benchmark for prediction accuracy and compared the performance with multiple classifiers in the ensembled form [13]. The paper also reports the variation in accuracy by modifying training configuration parameters like the number of epochs, network architecture, and the number of classifiers for ensembling. The analysis is performed on the publicly available Australian, German and Japanese credit datasets.

3.2 Market Risk

Market risk deals with the volatility of indices. Value at Risk (VAR) estimates the worst loss that will not exceed in a particular time frame. It captures the combined effect of underlying volatility and exposure to financial risks. Volatility estimation is a part of the time series forecasting domain. ANNs excel at these learning problems compared to traditional methods. A commonly used forecasting framework for time series is the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model [15]. It is derived from the Autoregressive Conditional Heteroscedasticity (ARCH) model, which predicts the volatility of innovation or error term as a function of previous error terms.

The volatility at time t is given as a function of prior error terms as:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{n} \alpha_i \epsilon_{t-i}^2 \tag{13}$$

GARCH model also considers the volatility of prior time to predict the current volatility:

$$\sigma_t^2 = \omega + \sum_{i=1}^{n} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 \tag{14}$$

Zhang et al. proposed a model to calculate the VAR (volatility) in an index from the GARCH model using ANNs [16]. The paper also presents a comparative analysis with other models like just GARCH and SVM to predict volatility.

3.3 Liquidity Risk

Liquidity risk is the risk of not being equipped to raise capital when required. Trading of financial instruments and assets enable the raising of capital to banking institutions. Such funding is known as funding liquidity. When it is difficult to borrow or raise money, liquidity risk comes into play. For example, banks faced heavy losses in the 2008 crisis when assets were trading at heavily discounted rates due to the crash and the banks urgently needed funds to keep their businesses running.

Machine learning can solve several liquidity risk factors. ANNs can be used to estimate a risk measure. Approximation of general risk trends and discovery of principal risk determinants can be made using ANNs. BN can be used to predict the probability of a liquidity risk event happening. By using distributive estimation, the ANN and BN implementations could identify the most critical liquidity risk factors.

Tavana et al. used a combination of ANN and BN to determine liquidity risk (LCR and NSFR) from a real-world dataset of a U.S. bank focusing mainly on loans for eight consecutive years. The ANN was used to approximate the general trend of liquidity risk and find the influential factors. Further, the BN was used to find the most influential factor from the data filtered by the ANN [7].

3.4 Operational Risk

Operational Risk arises due to malfunctions in internal and external processes. Machine Learning is used in an operational context to detect such malfunctions. Fraud detection is a significant area of focus of operational risk management.

Money laundering is the malpractice of hiding illegal sources of income by layering lawful transactions on such sources. This is done by routing money through many complicated but legitimate transactions to conceal the actual source. These funds are further used to support criminal, war and terrorist activities.

Sudjianto et al. employed clustering algorithms to segment customers based on transaction

activity [17]. The work combines clustering and profiling to perform unsupervised peer group analysis (PGA) [18]. These segments are used for analyzing if the group of customers were involved in money laundering. Another challenge in identifying fraud is to flag suspicious transactions first and then perform an anti-money laundering investigation. The paper also presents classifiers like LR, SVM, CART, C4.5, C5.0 and RF for determining illegal transactions.

Villalobos et al. has done comparable work on money laundering detection using ANN, SVM and C5.0 based classifiers [19]. The experiments were performed on a Latin American financial institution dataset, and decision tree-based classifiers excelled at the task compared to other models.

Table 1: Literature Survey Summary

Task	Author	Title	Model	Performance Measure	Dataset	Features
Credit Risk	Bellotti et al. [8]	Support vector machines for credit scoring and discovery of significant features	SVM, LR, KNN, LDA	AUC	Credit card history of 25,000 customers over a three month period	34 features taken from each customer's original application and features extracted from a credit reference agency
Credit Risk	Tsai et al. [13]	Using neural network ensembles for bankruptcy prediction and credit scoring	MLP, Ensembled MLP	Accuracy, Type I and Type II errors	Australian, German and Japanese credit datasets	20 attributes including age, gender, amount and duration of credit and job status
Credit Risk	Wang et al. [9]	A new fuzzy support vector machine to evaluate credit risk	Fuzzy SVM	Type I and Type II and overall accuracy	FAME, Japanese credit dataset	12 features for FAME, 15 for the Japanese credit dataset

Task	Author	Title	Model	Performance Measure	Dataset	Features
Credit Risk	Angelini et al. [14]	A neural network approach for credit risk evaluation	ANN	Error	Italian small business dataset	15 features derived from the balance sheet of the small businesses
Market Risk	Zhang et al. [16]	Calculating Value-at-Risk for high-dimensional time series using a nonlinear random mapping model	GARCH, ANN	p-value	China Securities Index 300 from April 8, 2005, to May 28, 2015	Daily closing price
Liquidity Risk	Tavana et al. [7]	An Artificial Neural Network and Bayesian Network model for liquidity risk assessment in banking	ANN, BN	MSE	Dataset of a U.S. bank focusing mainly on loans from 2004 to 2011	Liquid assets of bank, Current liabilities, short and long term deposits, total assets
Operational Risk	Sudjianto et al. [17]	Statistical methods for fighting financial crimes	LR, SVM, DT, RF, PGA	Error	AML Dataset	Account number, transaction amount, date, description
Operational Risk	Villalobos et al. [19]	A statistical and machine learning model to detect money laundering: an application	ANN, SVM, C5.0	Accuracy, AUC	Latin American Financial Institution data	Customer nationality, type of entity, source of assets, reputation risk score, owners risk rating

3.5 Summary

This chapter presents a comprehensive survey of various types of machine learning and statistical techniques used to solve risk management problems in finance. An interesting observation is that certain types of machine learning models perform better at specific risk management tasks. This is reflected by the number of academic works published for such algorithm-problem pairs. For example, SVMs have been extensively used for credit scoring tasks while tree-based classifiers excel at fraud detection tasks. Nevertheless, ANN seems to have found applications in all types of problems explored and produced state of the art results in many banking problems.

4 Case Study in Credit Risk

This chapter reports and analyzes the performance of various machine learning models on the credit risk classification task. The publicly available German Credit Score dataset is used to train the following classifiers: LR, LDA, KNN, CART, C5.0, NB, RF, SVM and ANN.

4.1 German Credit Dataset

This dataset classifies customers as having good or bad credit potential by a set of attributes. It contains 1000 instances of data points and 20 attributes for each data point. These features capture information like:

- Status of bank accounts
- Purpose of loan
- Amount and duration of loan
- Age, gender and employment of the customer
- Credit history
- Mode of repayment

4.2 Exploratory Data Analysis (EDA)

4.2.1 Distribution of Good and Bad Credit

The dataset contains 700 examples for the ”Good Credit” rating and 300 examples for the ”Bad credit” rating. This creates an imbalance in the dataset, which can degrade classification quality. Random oversampling is done to rebalance the dataset before feeding it to the models. [20]

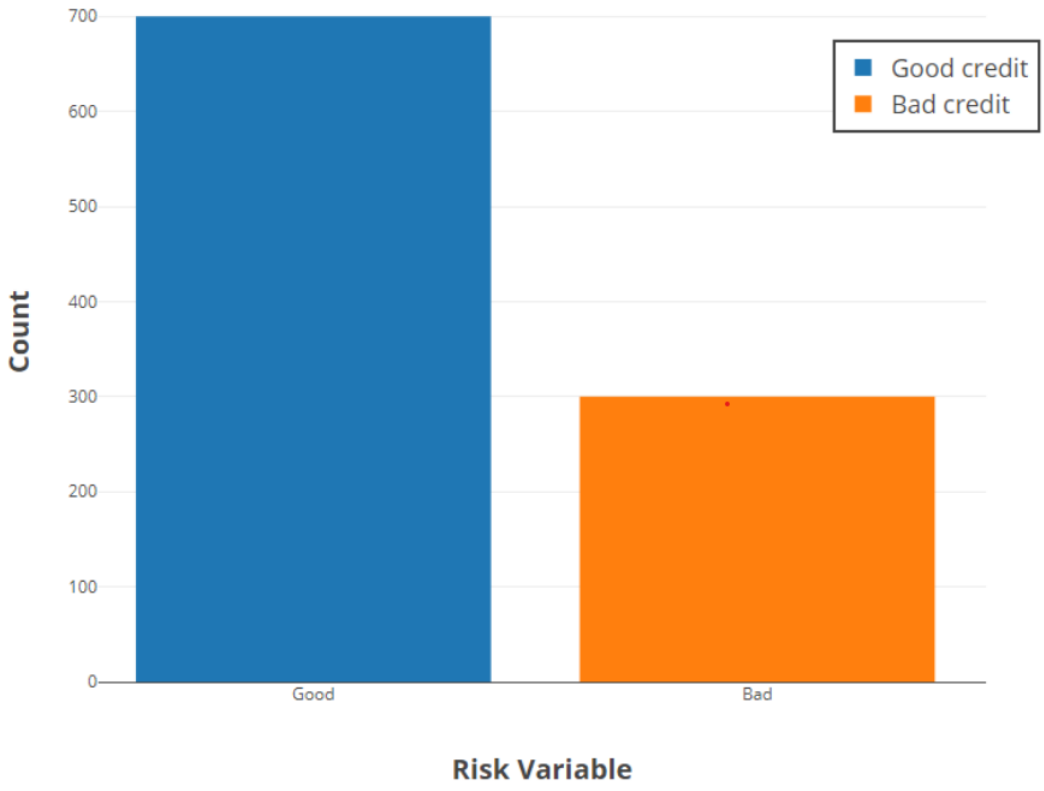

Fig. 8. Distribution of good and bad credit

4.2.2 Age Distribution

(Fig. 9) shows the distribution of loans with age. The distribution is shown based on the classification target as well as a combined graph. An interesting observation is that bad credit loans taper off quickly as age increases compared to the good credit loans, thus indicating that the younger audience may have a significant contribution in non-payment of loans.

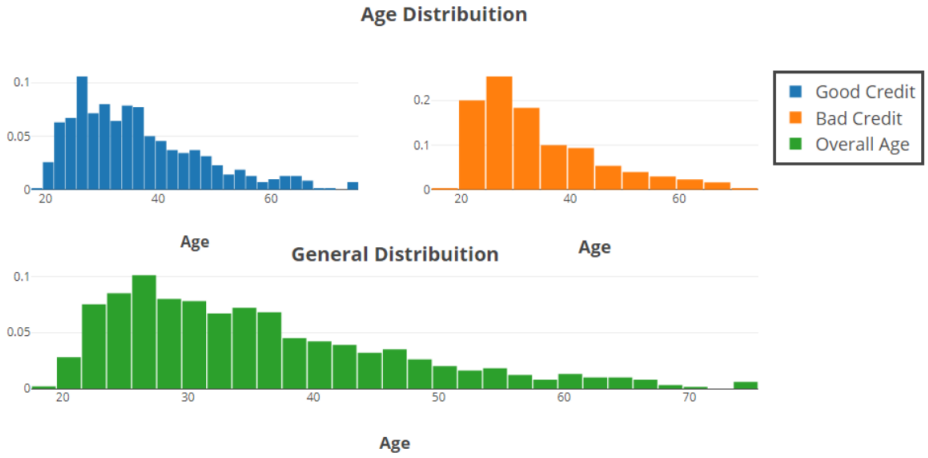

Fig. 9. Distribution of credit based on age

4.2.3 House Type Distribution

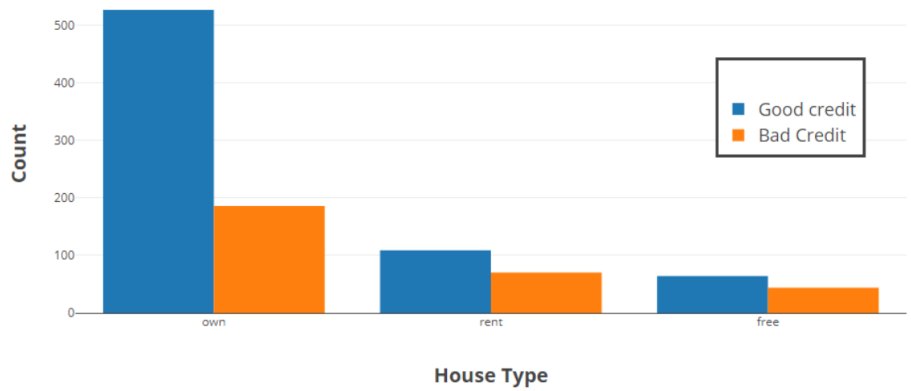

Fig. 10. Distribution of credit based on purpose

People who own their own house are more likely to pay off their loans. On the other hand, there is almost a 40% chance that the loan will have defaulted for the other two classes.

4.2.4 Loan Purpose Distribution

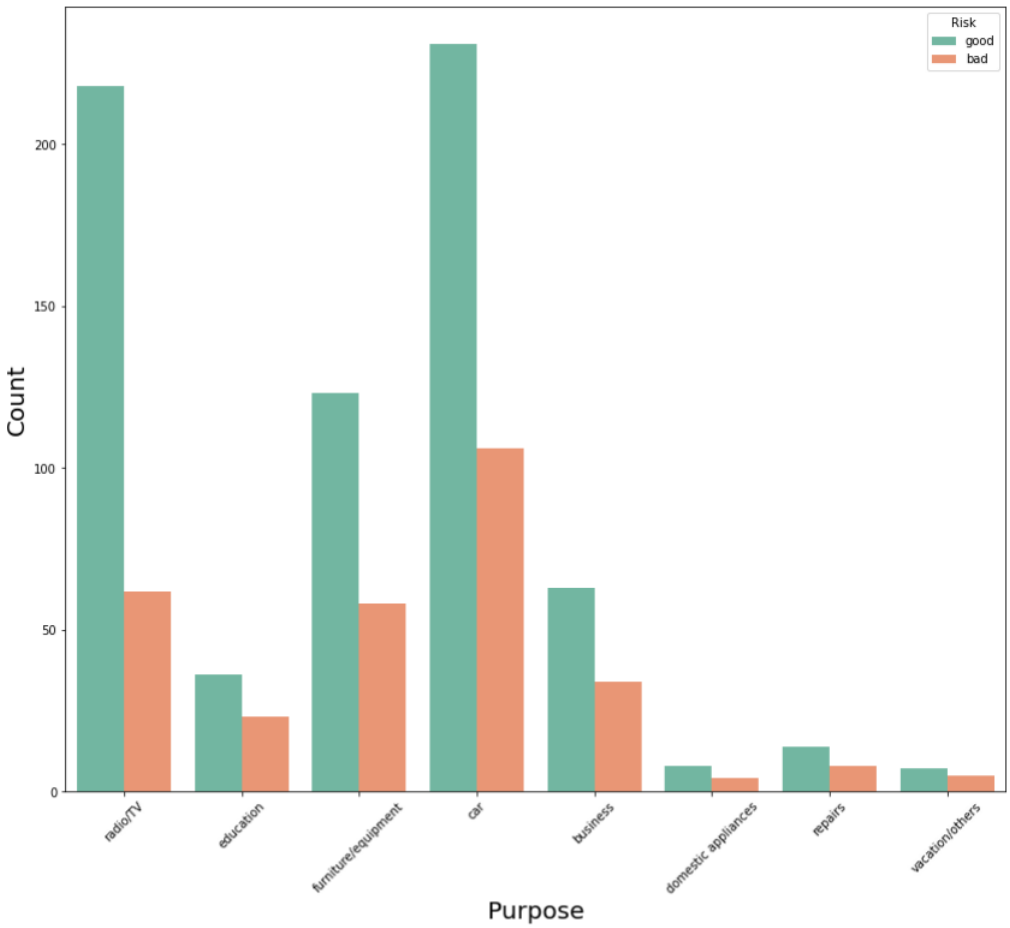

Fig. 11. Distribution of credit based on purpose

Loans are mainly taken to buy a car or TV. Credit taken to buy TV/Radio has the least risk of default. Education and Vacation loans carry a higher credit risk, which parallels with banks charging a higher rate of interest for education loans to account for the risk of default.

4.3 Results

Before training, the minority class was oversampled to meet the size of the majority class. The data was split into a 1:4 ratio with the larger portion used to train the models, and the smaller portion is used for evaluation. Five trials were carried out for each classifier, and their average performance metrics are reported in Table 2.

Table 2: Performance of classifiers on German Credit Dataset

Model	Accuracy	F1 Score	AUC
LR	0.7285	0.7300	0.7292
LDA	0.7188	0.7196	0.7196
KNN	0.7240	0.7406	0.7226
CART	0.8417	0.8529	0.8387
C5.0	0.8428	0.8557	0.8408
RF	0.8754	0.8820	0.8746
NB	0.7023	0.7112	0.7019
SVM	0.7955	0.8054	0.7959
ANN	0.7509	0.7586	0.7505

For each model, Receiver Operating Characteristics (ROC) curves are plotted in (Fig. 12).

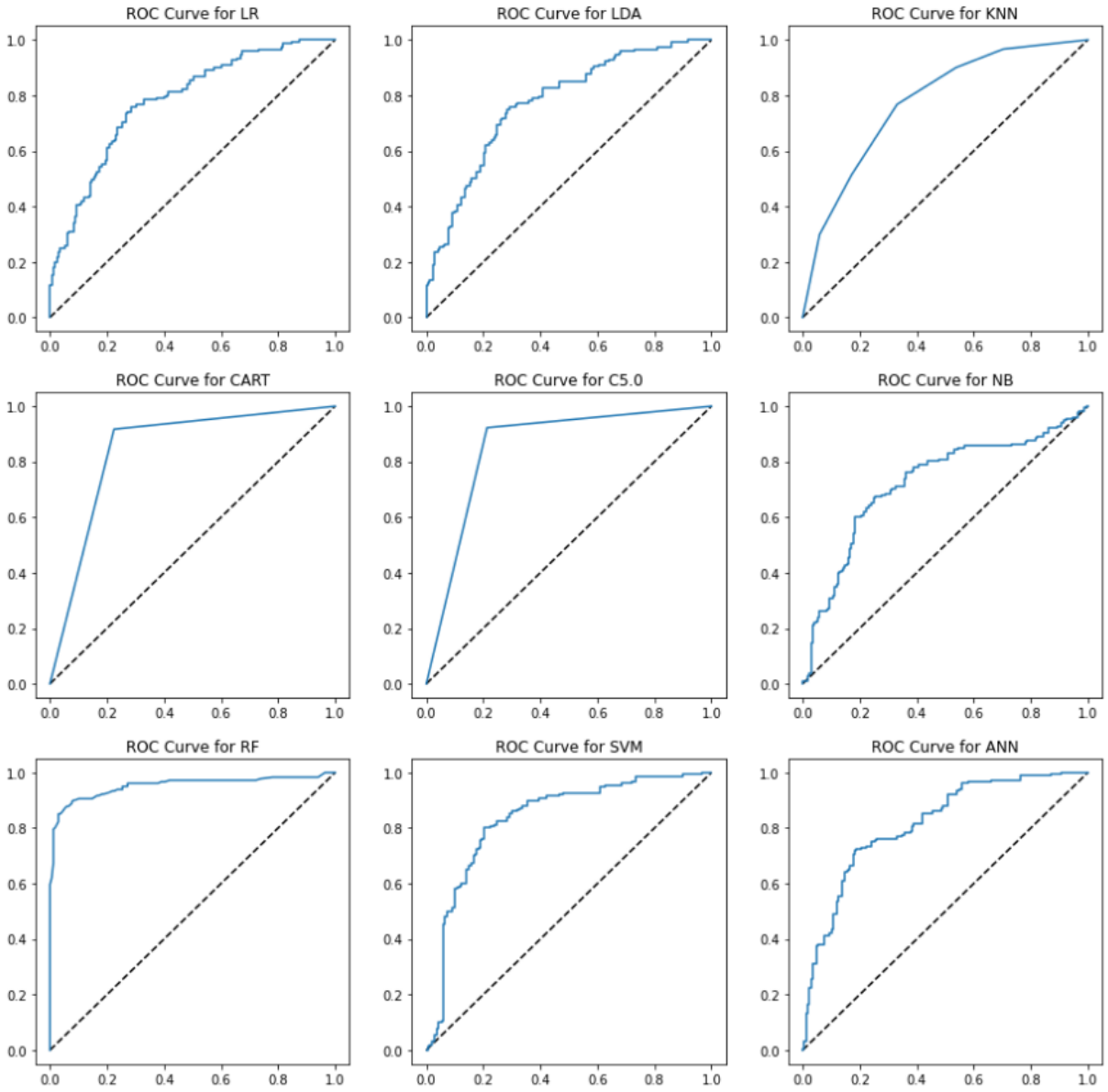

Fig. 12. ROC for classifiers

In summary, both the tree-based algorithms (CART and C5.0) performed better than other models with an F1 score greater than 0.8. When ensemble learning is used in the RF classifier, the best results are achieved with a 0.88 F1 score, which is reflected in the ROC curve. SVM and ANN are the next best classifiers for the learning task respectively.

5 Conclusion

Throughout all modern human history, banking has been an essential factor in fueling innovation and betterment of the society. With the coming of the information age, increased competition and modern banking facilities available to the commoner, risk management plays a vital role in building a sustainable and compliant banking business. Banks generate several terabytes of data that state-of-the-art machine learning models can leverage to make data-driven decisions. In this report, various machine learning applications for risk management in banking are presented. This work does a comprehensive survey of academic research for different risk management areas viz. credit risk, operational risk, market risk and liquidity risk. A case study is also presented comparing the performance of various classifiers like ANNs, DTs, SVM and others on a single dataset for credit scoring tasks.

References

[1] RBI. *Annual Report of the RBI for the Year 2020-21*. 2021. URL: https://www.rbi.org.in/Scripts/AnnualReportPublications.aspx?year=2021.

[2] BCC Research. *Technologies for Assessing Risk Management: Global Markets*. 2017. URL: https://www.marketresearch.com/BCC-Research-v374/Technologies-Assessing-Risk-Management-Global-10564858/.

[3] RBI. *Report on Trend and Progress of Banking in India, RBI, December 2020*. 2020. URL: https://rbi.org.in/scripts/AnnualPublications.aspx?head=Trend%5C%20and%5C%20Progress%5C%20of%5C%20Banking%5C%20in%5C%20India.

[4] Philippe Jorion. "Value at risk". In: (2000).

[5] Thirupathi Kanchu and M Manoj Kumar. "Risk management in banking sector–an empirical study". In: *International journal of marketing, financial services & management research* 2.2 (2013), pp. 145–153.

[6] Joel Bessis. *Risk management in banking*. John Wiley & Sons, 2011.

[7] Madjid Tavana et al. "An Artificial Neural Network and Bayesian Network model for liquidity risk assessment in banking". In: *Neurocomputing* 275 (2018), pp. 2525–2554.

[8] Tony Bellotti and Jonathan Crook. "Support vector machines for credit scoring and discovery of significant features". In: *Expert systems with applications* 36.2 (2009), pp. 3302–3308.

[9] Yongqiao Wang, Shouyang Wang, and Kin Keung Lai. "A new fuzzy support vector machine to evaluate credit risk". In: *IEEE Transactions on Fuzzy Systems* 13.6 (2005), pp. 820–831.

[10] Terry Harris. "Credit scoring using the clustered support vector machine". In: *Expert Systems with Applications* 42.2 (2015), pp. 741–750.

[11] Rong-Zhou Li, Su-Lin Pang, and Jian-Min Xu. "Neural network credit-risk evaluation model based on back-propagation algorithm". In: *Proceedings. International Conference on Machine Learning and Cybernetics*. Vol. 4. IEEE. 2002, pp. 1702–1706.

[12] Xin-yue Hu and Yong-li Tang. “Ann-Based Credit Risk Identificaion and Control for Commercial Banks”. In: *2006 International Conference on Machine Learning and Cybernetics*. IEEE. 2006, pp. 3110–3114.

[13] Chih-Fong Tsai and Jhen-Wei Wu. “Using neural network ensembles for bankruptcy prediction and credit scoring”. In: *Expert systems with applications* 34.4 (2008), pp. 2639–2649.

[14] Eliana Angelini, Giacomo Di Tollo, and Andrea Roli. “A neural network approach for credit risk evaluation”. In: *The quarterly review of economics and finance* 48.4 (2008), pp. 733–755.

[15] Tim Bollerslev. “Generalized autoregressive conditional heteroskedasticity”. In: *Journal of econometrics* 31.3 (1986), pp. 307–327.

[16] Heng-Guo Zhang et al. “Calculating Value-at-Risk for high-dimensional time series using a nonlinear random mapping model”. In: *Economic Modelling* 67 (2017), pp. 355–367.

[17] Agus Sudjianto et al. “Statistical methods for fighting financial crimes”. In: *Technometrics* 52.1 (2010), pp. 5–19.

[18] Richard J Bolton and David J Hand. “Peer group analysis–local anomaly detection in longitudinal data”. In: *Technical Report* (2001).

[19] Miguel Agustín Villalobos and Eliud Silva. “A statistical and machine learning model to detect money laundering: an application”. In: *Actuarial Sci. Dept. Anahuac Univ., Tech. Rep* (2017).

[20] Nathalie Japkowicz and Shaju Stephen. “The class imbalance problem: A systematic study”. In: *Intelligent data analysis* 6.5 (2002), pp. 429–449.

Acknowledgment

I would like to express my deep gratitude and indebtedness to my project guide, Dr Rupa G. Mehta, Associate Professor, Computer Engineering Department, SVNIT Surat, for her valuable guidance, constructive feedback and co-operation with a kind and encouraging attitude for the successful completion of this work. I would also like to thank Dr Mukesh A. Zaveri, Head of Department, Computer Engineering Department. I am also thankful to the Computer Engineering Department faculties for direct or indirect support to complete this seminar.