Innovations in Systems and Software Engineering
https://doi.org/10.1007/s11334-022-00522-x

S.I. : LOW RESOURCE MACHINE LEARNING ALGORITHMS (LR-MLA)

Automated credit assessment framework using ETL process and machine learning

Neepa Biswas[1] · Anindita Sarkar Mondal[1] · Ari Kusumastuti[2] · Swati Saha[3] · Kartick Chandra Mondal[1]

Received: 29 January 2022 / Accepted: 9 December 2022

Abstract
In the current business scenario, real-time analysis of enterprise data through Business Intelligence (BI) is crucial for supporting operational activities and taking any strategic decision. The automated ETL (extraction, transformation, and load) process ensures data ingestion into the data warehouse in near real-time, and insights are generated through the BI process based on real-time data. In this paper, we have concentrated on automated credit risk assessment in the financial domain based on the machine learning approach. The machine learning-based classification techniques can furnish a self-regulating process to categorize data. Establishing an automated credit decision-making system helps the lending institution to manage the risks, increase operational efficiency and comply with regulators. In this paper, an empirical approach is taken for credit risk assessment using logistic regression and neural network classification method in compliance with Basel II standards. Here, Basel II standards are adopted to calculate the expected loss. The required data integration for building machine learning models is done through an automated ETL process. We have concluded this research work by evaluating this new methodology for credit risk assessment.

Keywords Data integration · ETL · Data warehouse · Machine learning · Automated credit risk assessment

1 Introduction

In today's world, data are the most vital part of an enterprise. Data of an enterprise are spread across different heterogeneous data sources. Data from different sources are consolidated in a data warehouse (DW) through the ETL (extraction, transformation, and load) process. The data from different sources in the ingestion layer are pulled in batches resulting in a loss of real-time insights and revenue opportunities. Business analytics based on real-time data is crucial to take strategic decisions or supporting operational activities. Enterprise should have a process to integrate data from different source systems in near real-time [17,18] and replicate the data in DW. Automated ETL process [43] needs to be established in the enterprise to capture the data changes in source systems, place the changed data in the staging area, perform the required transformation and finally ingest data in DW in real-time.

Here in this paper, we will discuss a specific use case in the Banking and Financial domain—automated machine learning-based credit risk assessment. The likelihood that a borrower would not repay their loan to the lender is called credit risk. The lender should assess the credit risk of each borrower as precisely as possible. The inability or failure to estimate borrowers' probability of default can have major consequences for lenders as well as for the national economy. The main reason that led to the global financial crisis in 2008

Neepa Biswas, Anindita Sarkar Mondal, Ari Kusumastuti and Swati Saha have contributed equally to this work.

✉ Kartick Chandra Mondal
kartick.mondal@jadavpuruniversity.in

Neepa Biswas
biswas.neepa@gmail.com

Anindita Sarkar Mondal
sarkar.anindita5@gmail.com

Ari Kusumastuti
arikusumastuti@gmail.com

Swati Saha
swati.saha@tcs.com

[1] Department of Information Technology, Jadavpur University, Salt Lake Campus, Kolkata, West Bengal 700106, India

[2] Department of Mathematics, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Malang, Jakarta, Indonesia

[3] Tata Consultancy Services Limited, Kolkata, West Bengal 700156, India

Published online: 31 December 2022

was the high default rates of home mortgages in the USA. There is a certain amount of credit risk [27,37] associated with every borrower.

The expected loss of a given loan is calculated as the Probability of Default (PD) multiplied by both the Loss Given Default (LGD) and the Exposure at Default (EAD) [28]. PD is the likelihood that a borrower would not be able to pay their debt. In other words, it is an estimate of the likelihood that the borrower would default. LGD is the share of the loan amount that is lost if the borrower defaults; it is the proportion of the total exposure that cannot be recovered by the lender. EAD is the total loss in terms of the amount the lending institution is exposed to.

As per Basel II standards [13,35], either a standard approach or an internal rating-based (IRB) approach [61] can be taken for forecasting the expected loss. Here the internal rating-based approach is considered where the bank is allowed to build its own models for different components of expected loss. In foundation internal rating-based approach (F-IRB) [35], only the probability of default model is built by the bank. But in the advanced rating-based approach, all three risk components PD, LGD, and EAD are estimated internally within the bank instead of relying upon any external party/agency. The advantage of the advanced IRB approach is that bank can use their own estimated parameters for assessing credit risk.

This credit assessment framework can be used by lending institutes as a credit decision-making tool that can accelerate their future growth and competitiveness. This can also be used to monitor existing loans by reassessing credit risk. The disruptions created by Covid 19 pandemic lead banks and financial institutes to build their own credit risk assessment tool based on real-time data. Considering the serious implication of Covid 19, the lending institution should also reevaluate the credit risk. Customer credit demand, as well as credit requirements for small and medium businesses, has increased after the worldwide lockdown of the Covid 19 pandemic. Using this credit assessment framework will reduce the turnaround time for credit decisions significantly. Lenders can respond to customers very quickly.

The global financial market has several types of risks. Many research approaches have been made to identify an efficient way to predict future risk. This paper aims to work on the credit risk management of any financial domain in compliance with Basel II Standards. In this paper, machine learning-based credit risk assessment is done over the data uploaded in the data warehouse through an automated ETL process. An architecture is designed to build a credit assessment framework using an automated ETL process using ML-based solutions. A logistic regression (LR)-based supervised learning approach and neural network (NN)-based method have been employed here for building the model. Logistic regression, decision tree, support vector machine, random forest, and extreme gradient boosting methods are the commonly used model that has been used in the past for credit risk analysis. This proposal aims to build an automated data integration system for evaluating credit risk. Here automation is applicable at every phase of ETL, like automated data extraction, cleaning, and loading processes that were proposed in our previous research article [43]. The novelty of this paper is that, along with credit scoring in the financial domain, we are also evaluating an automated data integration system.

Recently, some research has been initiated on credit risk models using machine learning algorithms. But none of the approaches have followed Basel II Standards [49]. We are following the global standard of the Basel Committee on Banking Supervision (BCBS) which is a more realistic approach to addressing credit risk in the banking sector [13]. For credit risk assessment, evaluation is done using supervised machine learning algorithms as well as Neural networks. Implementation is done on U.S.-based Lending Club loan data set. In future work, various classification algorithms can be evaluated.

In this paper, we will focus on the end-to-end process to estimate Expected Loss which is associated with credit risk using an automated ETL process and ML-based model. The process starts with capturing the borrower's data and other required data from source systems, performing required transformations, and doing continuous ingestion in DW.

The paper is organized in the following way. Section 2 briefly discusses some notable related work in credit risk assessment, ML-based credit risk, and ETL automation domain. The standard credit risk modeling approach is discussed in Sect. 3. Section 4 contains information related to data which covers data sources, data warehouse design, and dependent variables for the PD model. Solution approaches of automated credit assessment framework are discussed and corresponding architecture is built in Sect. 5. A detailed discussion about the implementation of ML models is included here. Finally, Sect. 6 concludes the work with brief summary followed by targeted future scope.

2 Related work

The main focus of this research is to build ML-based credit risk modeling using real-time data which is loaded in DW through an automated ETL process. Some research progress related to the ETL process and credit risk modeling is discussed in this section. Credit risk modeling is an elementary process of banking and had been introduced long back. In 1968, the first multivariate credit model specification is given by Altman [7] which is known as Altman's-Z score model. Apart from that, other authors also have estimated different types of risk models introduced by Frydman et al. [33], Li

[41] and Shumway [51]. Broadly credit risk modeling is categorized into four groups [24]. The first group is based on 'Merton' structural approach for valuing risky debt. According to this, the firms default if the total asset of the firm is less than the total liabilities. The second group is the economic factor risk model where risk is determined by several macroeconomic factors. For both models, the risk is computed at the individual entity level. Top-down models like credit risk plus are included in the third group. The last group consists of models which use non-parametric methods. Carey majorly worked on these non-parametric methods [19]. In the last two decades, commercial banks show their interest to develop internal credit risk modeling to better assess their risk. Internal rating-based (IRB) approach had been taken by banks starting from mid-90 and especially Basel II regulation first introduced in 2004 [26]. There are some models which are specifically designed for small and medium enterprises by Edmister [29], Altman and Sabato [8], Altman et al. [9], and most recently by Altman et al. [6] and Soui et al. [55].

In the early period of credit analysis, logistic regression [32], probit regression [36], and discriminate analysis [39]-based traditional statistical models have been used. Afterward, for the intensive computational demand, artificial intelligence (AI)-based approaches like rough set (RS) [11], support vector machine (SVM) [42], decision trees (DT) [21], and neural network (NN) [22] have been introduced. Because of the computational efficiency recently, some research has adopted machine learning-based approaches like LR [10], SVM [12], NN [50], random forest [59], or combination strategy [40] in the credit risk field. A comparative review of different credit risk analysis approaches is presented in Table 1.

Regarding ETL processing, many conceptual ETL modeling approaches have been developed in recent years. These conceptual modeling patterns can be categorized as UML language-based [44,58], meta model-based [30,63], BPMN language-based [3,5,47], semantic web technology-based [52,53], and SysMl language-based [15,16] approach. An MDA (model-driven architecture)-based approach [45] has been proposed for designing ETL model which enables automatic code generation from the conceptual model. Article [52] describes a semantic web-based ETL design with a high level of automation. Embley et al. [31] proposed an ontology-based conceptual model for automatic data extraction. Article [4] proposed a model-driven ETL framework using BPMN language. The model-to-text transformation technique proposed in this work can automatically produce executable code. This type of code is compatible with any commercial ETL tool. Moreover, the model-to-model transformations procedure is able to automatic code updation for maintenance purposes. Automatic data loading into the DW [20] is proposed by tracking any business events from any application. An architecture designed by Suresh et. al. [56] will automatically optimize ETL throughput. Radhakrishna et. al. have come up with an idea of an automated ETL process by the use of scripting technology [48]. The three ETL jobs (extraction, transformation, and load) can be evaluated by any ETL tool using scripting technology.

Regarding the real-time ETL process, various technical challenges and possible solutions was first discussed by Vassiliadis et al. in [62]. For continuous data integration, an efficient methodology is discussed in article [38] to perform continuous data loading process. A log-based change data capture (CDC) methodology is projected by H. Zhou et al. in [64]. A triggering and scheduling-based ETL framework has been designed in article [54] for real-time data refreshment in the DW. For real-time ETL processing, an incremental loading approach has been implemented by the snapshot-based CDC approach in article [18]. Although some research work has been found for addressing real-time ETL and automated ETL processing. It is comparatively a new domain of research. So, credit risk modeling and real-time ETL processing, both of these issues are gaining popularity in recent times as well as it is still an open problem.

3 Credit risk modeling approach

Bankruptcies of big financial institutions led to huge disturbances in the economy, and millions of people experience significant financial difficulties. To prevent such consequences, regulators imposed certain requirements on banks to make sure that banks can carry out their business without risking the stability of the economic system. In this proposal for credit risk assessment, we have followed the Basel II standards.

3.1 Credit assessment based on Basel II accords

The Basel II is an international business standard [13] defined by Basel Committee on Banking Supervision regarding risk and capital management requirements to ensure that the bank has adequate capital to guard against the risks that the bank exposes itself through its lending investments. Capital allocation of banks needs to be more risk-sensitive. The higher the risk the bank is exposed to, the more capital the bank needs to hold for overall economic stability.

The first pillar of the Basel II accord called minimum capital requirements deals with the major types of risk of bank faces credit, operational, and market risk. As per Basel II standards, as shown in Fig. 1, there are two different approaches to model credit risk—standardized approach (SA) and internal rating-based approach (IRB). According to Basel II, banks can choose any one approach for modeling credit risk or calculating expected loss. In a standardized approach [60], banks use data from external credit agencies

Table 1 A comparative review of different credit risk analysis

Year & Ref.	Algorithm	Data set	Features
2022 [59]	Random Forest	Chinese HD data set	Small business data
2017 [10]	LR & NN	Small data sample	No noise reduction
2017 [42]	SVM	Chilean bank data	Combination of data sources
2019 [11]	Rough Set	Chinese bank data	Fuzzy RS, Fuzzy C-means clustering
2021 [21]	Decision Tree	Car loan data in Taiwan	Rule-based approach
2016 [22]	ANN	Hybrid models uses	Evaluated on five data sets
2018 [32]	Hybrid LR	Production data in Iran	Delphi method used
2019 [50]	Ensemble model	German credit data set	Personal Credit Risk Assessment

to assess the credit risk of borrowers. For example in the USA, Fitch Ratings, S&P, and Moody's are popular Credit Rating Agencies. In India, TransUnion Credit Information Bureau (India) Limited (CIBIL) gives the credit score rating generally named CIBIL score that is used for the same purpose. The standardized approach completely relies on ratings given by external agencies. External credit ratings are reliable and assess creditworthiness well. However, banks are collecting a lot of data in the process of taking loan applications and gaining additional information about the behavior and operations of entities after they have been granted loans. Detailed credit estimation can be done using this additional information.

The Basel II accord allows banks to use this additional information to calculate their internal risk rating under the IRB approaches. This approach [61] can be either foundation level or advanced. In any of the IRB approaches, expected loss is calculated as the product of PD, LGD, and EAD. The only difference between these two approaches is in the components that banks can estimate their own. In foundation internal rating-based approach, PD is modeled by the bank. LGD and EAD are provided by regulators. Under the advanced IRB approach, the bank estimates all three components by itself.

The proposed solution intends to build different ML models to predict each of these components (PD, LGD, and EAD) and then compute them to obtain the expected loss for a given exposure level.

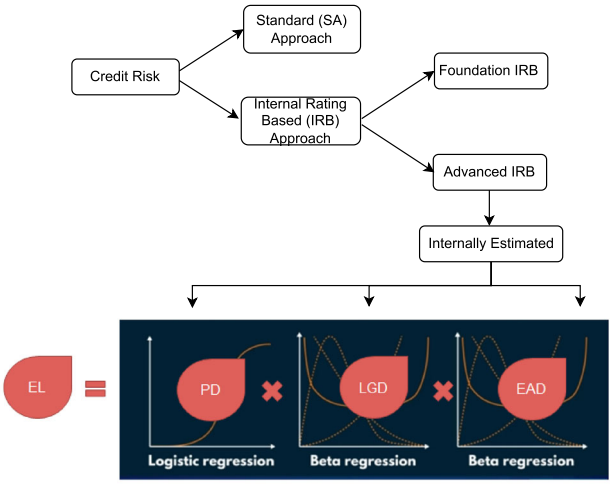

Fig. 1 Credit risk approach

3.2 Methodology for credit risk modeling

Traditionally lending institutions calculate expected loss manually to assess risk. Manual credit risk assessment takes a lot of time, also it is error prune. The manual process of expected loss calculation is not very consistent as there is no single/source of truth. Also, in the traditional process, creditworthiness is mainly determined by the borrower's credit history. In the last decade, an expert-based credit scoring model was introduced to determine whether borrowers can fulfill their requirements. Now with the technological advances, most lending institutions want to update their credit assessment process due to stronger regulation and better risk management. The aim of ML modeling is to collect and analyze the data to develop an appropriate model. ML-based credit assessment measures the credit risk of applicants much more accurately and quickly than the traditional process [1,14]. It considers data from different sources, i.e., various borrowers' information, data from previous loans, repayment behavior, and data from external agencies to predict credit scores instead of relying only on borrowers' credit history. Traditional models consider a linear relationship between credit score and the data whereas ML models can capture the complex non-linear relationship that is present in data. Hence, predicting power is more in ML-based models compared to traditional processes [25,34]. It helps enterprise to make a smarter decision and responds to competitors and market change. ML-based credit assessment covers many benefits like reducing time to credit decisions, improving accuracy, proactive risk management, increasing efficiency,

cost optimization, assessing credit risk without credit history and many more.

4 Data sources

Data from different source systems are aggregated in DW. Lenders can leverage historical data to predict the likelihood of a loan default. The data set considered for this work is provided by a large US loan giving company Lending Club [1] which contains consumer loans issued from 2007 to 2018 [2].

4.1 Data warehouse design

Star schema-based DW is designed for the proposed model as shown in Fig. 2. Here fact_loan is a fact table that has references to different dimension tables. Fact table has following variables, viz. loan_id, member_id, funded_amount, loan_amount, int_rate, installment, loan_status, issue_date, total_pymnt, total_pymnt_inv, total_int_rcv, total_late_fee _rcv, total_prncp_rcv, out_prncp, out_prncp_inv and recoveries. Fact table fact_loan also has references to the customer dimension table (dim_customer), loan application dimension table (dim_loan_application), and date dimension table (dim_date). The customer dimension table (dim_customer) contains customer-specific information. It has reference to the location and date dimension table (i.e., dim_location and dim_date). The location dimension table (dim_location) contains the address of the customer and has the following attributes zip_code and addr_state. The date dimension table contains information of a date like month, year, and quarter details. Another dimension table (dim_loan_application) contains information related to loan applications having the following attributes pymnt_plan, title, purpose, application_type.

4.2 Definition of default

We need to classify the loan as a good or bad loan. At first, we need to define the 'Default' Typically 'Default' is defined based on the delinquency of the borrower measured in days past the payment due date. Generally, a loan is considered defaulted if payment is due for more than 90 days. In the selected data set, there is only one column Loan Status about borrowers' performance and loan status. Each loan has one of these statuses: Charged Off, Current, Default, Does not meet the credit policy Status: Fully Paid, In Grace Period, Late (16–30 days), Late (31–120 days). Evidently, accounts

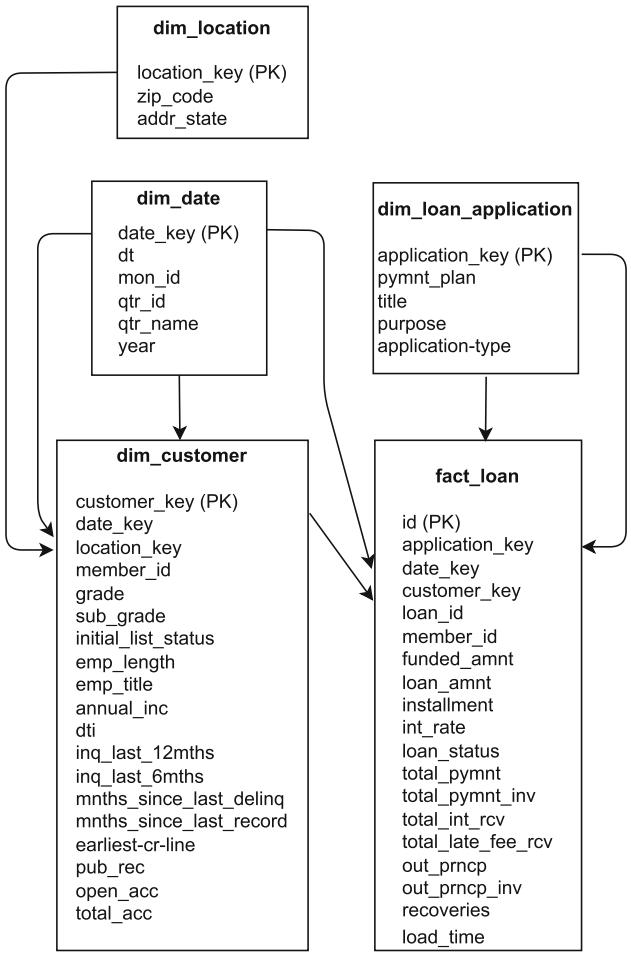

Fig. 2 Data warehouse schema

that have been fully paid have not defaulted, therefore they are good. On the other hand, Charged Off, Default accounts have definitely defaulted, therefore they are bad. The loan status is considered a bad loan if the loan status has any of these values: Does not meet the credit policy- Status: Charged Off, Late (31–120 days). The Default/Non-default indicator is stored in a new variable called the Good–Bad indicator.

4.3 Dependent & independent variables For ML model

Three regression models are going to build the Probability of Default (PD) model, a Loss Given Default (LGD) model, and an Exposure At Default (EAD) model. For the PD model, logistic regression is used, while for the LGD and EAD models, beta regression is used. These models are used to predict our outcome of interest also called dependent variables. The variables which are used to predict the dependent variable are called independent variables, predictors, or features. The Good–Bad indicator indicates whether the borrower is defaulted or not. LGD model calculates how much of the loan

[1] https://www.lendingclub.com/investing/peer-to-peer.

[2] https://www.kaggle.com/wordsforthewise/lending-club.

was recovered after the borrower has defaulted. This information is contained in the recoveries column. EAD model calculates the total exposure at the moment the borrower compared to the total exposure in the past. This information is present in the total recovered principal column.

5 Proposed machine learning-based solution

An architecture is designed to build a credit assessment framework using an automated ETL process based on ML. Figure 3 shows the overall proposed architectural design of the approach.

The main objective of this proposal is to build an automated data integration system. To automate the ETL process a Continuous Integration (CI) [23] platform is evaluated. For Continuous Integration, an open-source automation server Jenkins [3] is used. Jenkins is integrated with the Data Integration tool Informatica. It can promote building, testing, deploying, and releasing database changes in a faster and more frequent way. The Jenkins pipeline will execute automated scripts to process the ETL steps.

Data pre-processing step is crucial as far as data quality is concerned. The success of the ML model largely depends on the quality of the data. In this section, general techniques for data pre-processing are discussed. The pre-processing steps of discrete and continuous variables are summarized in Fig. 4.

5.1 Pre-processing techniques

5.1.1 Fine classing

It is the process of grouping variables into some initial categories. This process is required for continuous variables. For example, consider a variable "month since issue date" which has around 100 distinct values. This variable needs to be divided into some initial categories.

5.1.2 Coarse classing

It is the process of constructing new categories based on the initial ones. Categories that have a similar weight of evidence are combined into bigger categories. Using course classing, the number of dummies is reduced.

5.1.3 Weight of evidence

It shows the extent to which each of the different categories of an independent variable explains the dependent one. In other words, WOE represents how much evidence the independent variable has with respect to differences in the dependent variable. Here is the formula for calculating WOE:

$$\text{WOE} = \ln(\text{Event\%}/\text{NonEvent\%})$$

For the PD model, outcomes can be of two types: Non-defaulted and defaulted. So, the weight of evidence would be the natural logarithm of the ratio of the percentage of non-defaulted in a particular group from the total number of defaulted that falls into the category.

Table 2 Interpretation of information values

Information value (IVal)	Predictive power
$< .02$	Not useful for prediction
$.02 - .1$	Weak predicting power
$.1 - .3$	Medium predicting power
$.3 - .5$	Strong predicting power
$> .5$	Too Strong to be true

5.1.4 Information value

It is used to identify the independent variables which explain the dependent variable best. The formula for calculating information value is,

$$\text{IVal} = \sum_{n} (\text{Event\%}/\text{NonEvent\%}) * \text{WOE}$$

IVal helps us select the predictors and variables that we choose for the ML model. It is always in the range between 0 and 1 and how the information values are interpreted is shown in Table 2. We have calculated the information value for all the variables to assess their predicting power.

5.1.5 Pre-processing of discrete variables

For some variables, the value contains unnecessary text which needs to be removed. For example, for variables emp_length and term, clean-up is done by removing unnecessary text and converting them to float type. Dummy variables are created for discrete variables, e.g., purpose of the loan, home ownership, grade, sub-grade, verification status, state, etc. If there are too many categories or two similar categories are present, several dummies are bundled up into one based on similar WOE. The weight of evidence (WOE) of different variables is examined to check if any grouping of categories is required or not. When these dummy variables are put into a regression model, one category for each variable needs to be kept out against which the impact of all others on the outcome will be assessed.

[3] https://jenkins.io/.

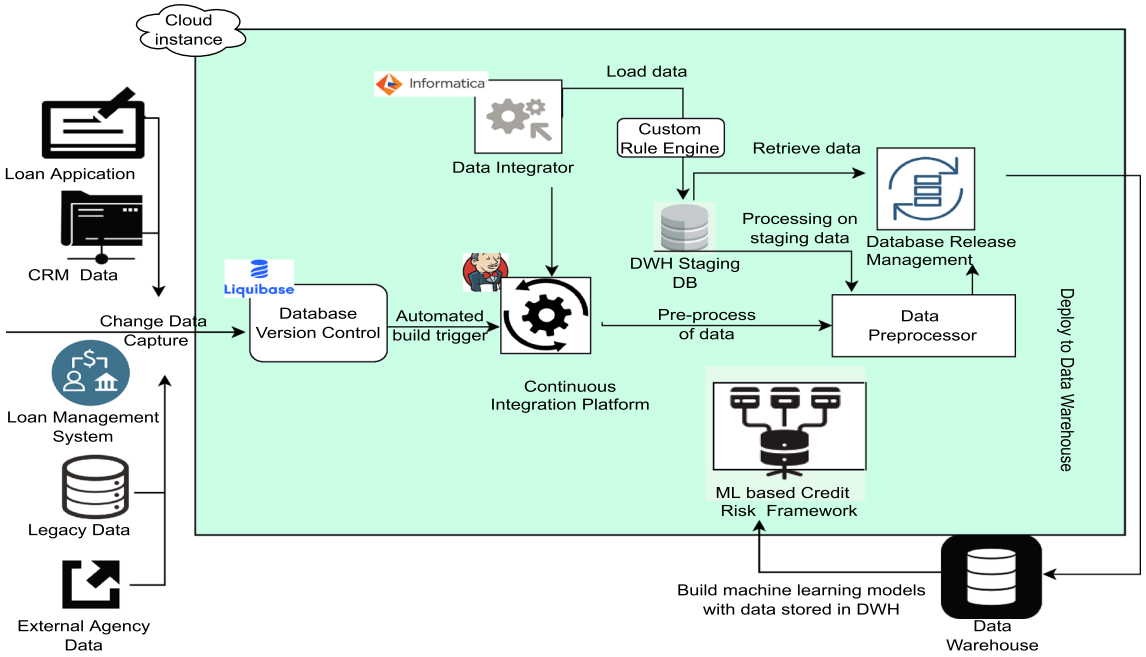

Fig. 3 Proposed architecture diagram for credit risk analysis

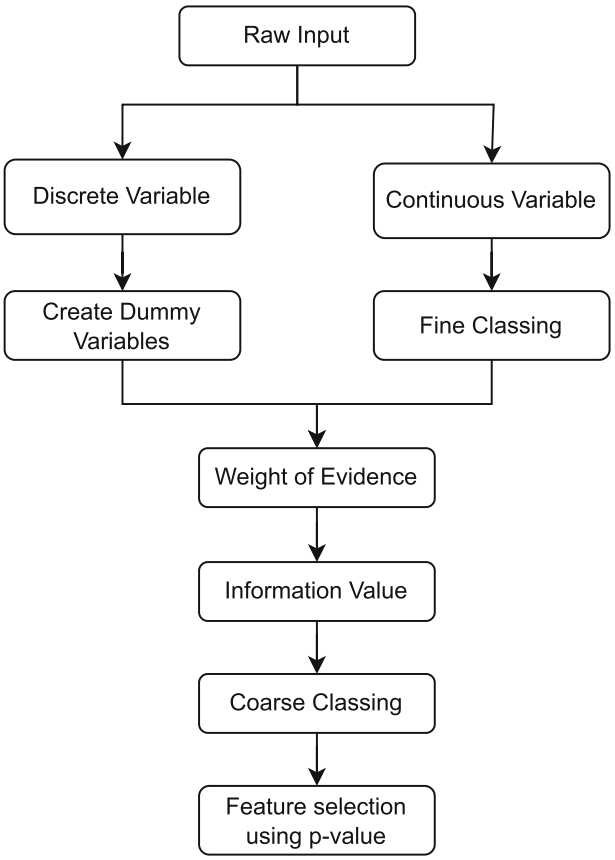

Fig. 4 Data pre-processing steps

5.1.6 Pre-processing of continuous variables

To deal with the missing values of the attributes, the total revolving credit limit is filled with the total funded amount, or for annual income, the missing value is replaced by the mean value of existing (non-missing) values. For some variables, missing values are filled with zeros like month_since_earliest_cr_line, acc_now_delinq, total_acc, pub_rec, open_acc, inq_last_6mnths, delinq_2years. A new variable corresponding to each date variable is computed which is basically the difference between the current date and the value of the date variable. Continuous variables can take any value in a given range. Hence it is difficult to convert continuous variables to dummy variables. Let's consider continuous variable months since the issue date. This attribute is derived from the issue date of the loans available in the original data set. This attribute has about 100 distinct values. We need around 50 initial categories so that course classing can be done neatly. First continuous variables are turned into the number of initial categories of equally sized intervals which is called Fine Classing. Once the fine classing is done, we can treat this variable just like any other categorical variable. Once continuous variables are grouped into initial categories, we explore how well each of them discriminates between defaulted and non-defaulted loans. If two adjacent categories discriminate equally well, those categories are merged. If the next category discriminates a lot better or a lot worse than the previous one, it is treated as a separate category. The weight of evidence is calculated for each of the categories. Then consecutive categories with similar WOE are grouped

together. This process is called coarse classing. Once continuous variables are reached the final version of categorize, then dummy variables are created for the new category.

5.2 Probability of default model

PD model [24] is a logistic regression model with a binary indicator for good or bad as a dependent variable and only dummy variables as independent variables. Logistic regression estimates the relationship between a dependent variable and independent variables. The logistic regression model predicts the probability of an event occurring. It predicts binary outcomes as defined by a logistic function. Let's consider one independent variable, Annual Income. Majorly borrowers having lower annual income have defaulted. Hence, the greater the annual income, the lower the probability of default.

Interpret ability is extremely important for the PD model as it is required by regulators. PD model should be very easy to understand and apply. The established practice for the PD model is that all independent variables need to be dummy variables so that even people who do not have any knowledge of statistical analysis, also should be able to work with it.

In logical regression, the first relationship between the dependent variable and the independent variables is assessed. Then regression coefficients of each of the independent variables are estimated. The positive coefficient of independent variables indicates positive quality which means higher creditworthiness. If we consider "1" in the Good_bad flag to represent default and "0" to represent non-default account, by executing the regression model, we got positive coefficients for the debt-to-income ratio. It indicates the probability of default increase with debt to income ratio. Generally in real life, we prefer to associate positively with positive feelings. Positive coefficients would signify a better quality/state of the borrower's account. Here the scenario is exactly the opposite. Hence we set "1" in the Good_bad flag to represent non-default and "0" to represent the default account. Logistic Regression is similar to Linear Regression.

Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m$$

Logistic Regression

$$P(Y=1) = e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_m X_m)} / (1 + e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_m X_m)})$$

$$P(Y=1) = e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_m X_m)} / (1 + e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_m X_m)})$$

$$P(Y=1)/(1-P(Y=1)) = e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_m X_m)}$$

$$P(Y=1)/(P(Y=0)) = e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_m X_m)} - - > \text{Odds}$$

$$ln(P(Y=1)/(P(Y=0))) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m$$

Calculate regression coefficients

In PD Model, all the independent variables are dummy variables. So values of the independent variables are either 1 or 0.

If $X_1 = 1$,

$$ln(P(Y=1)|(X_1=1)/(P(Y=0)|(X_1=1))) = \beta_0 + \beta_1 + \beta_2 X_2 + \cdots + \beta_m X_m$$

If $X_1 = 0$,

$$ln(P(Y=1)|X_1=0/(P(Y=0|X_1=0))) = \beta_0 + \beta_2 X_2 + \cdots + \beta_m X_m$$

$$ln(P(Y=1)|(X_1=1)/(P(Y=0)|(X_1=1))) - \ln(P(Y=1)|X_1=0/(P(Y=0|X_1=0))) = \beta_1$$

$$ln(\text{Odds}(Y=1|(X_1=1)/\text{Odds}(Y=1|X_1=0)) = \beta_1$$

$$\text{Odds}(Y=1|(X_1=1)/\text{Odds}(Y=1|X_1=0) = e^{\beta 1}$$

Thus the ratio of odds of an event occurring for observation with the dummy variable having a value of 1 to the observation with the dummy variable having a value of 0 equals in exponents to the power of the regression coefficient of the dummy variable.

5.2.1 PD model building using logistic regression with *p* values

PD model is built using logical regression. A logical regression model from sklearn is imported. An instance of the logical regression class (rag) is being created. PD model is estimated by fitting the inputs and targets to do that. The fit method of the 'rag' object is used which takes two data frames: the input data frame and the target data frame. The input data frame contains all the dummy variables which need to be included in the PD model. Loan data targets data frame contains dependent variable Good–Bad flag. Once the coefficients of dummy variables are getting calculated, a summary table is getting created which contains feature names and their corresponding coefficients.

As of now, all independent variables are used to create a PD model which is uniquely varied. That means, the impact of each independent variable is taken into account individually, and the collective impacts of all the features are not considered. Now, only the independent variables which contribute to predicting the default of borrowers are added to the

model. The '*p* value' method is used to check the statistical significance of the coefficients of each dummy variable. In the '*p* value' method, the impact of all the features on the outcome is collective rather than independent. The logistic regression model does not have a built-in way to calculate these multivariate *p* values. We can achieve this to alter the fit method from the logistic regression module.

From the Logistic Regression with *p*-value class, *p* values of the coefficients of the independent variables can be extracted using the p method. Then the summary table is created with an additional column *p*-value. Now, we can select independent variables based on p values by retaining the variables with coefficients that are statistically significant.

One independent variable is represented by one or more dummy variables. So, if the coefficients for the dummy variables correspond to an independent variable and are all statistically significant, then all dummy variables corresponding to that independent variable need to be retained. If none of them are statistically significant, those variables need to be removed. If one or a few dummy variables represent an independent variable, then all dummy variables corresponding to that independent variable are retained. Generally, if the *p*-value is less than 0.05, then the variable is considered significant.

The coefficients for all dummy variables that represent the grade variable are statistically significant, thus dummy variables corresponding to "Grade" needs to be retained. The coefficients for all dummy variables that represent the Home Ownership variable are also statistically significant. So, we keep these dummy variables as well. For verification status, as the coefficients of one dummy variable are statistically significant, this variable is retained. In the case of the address state, all dummy variables are significant except the first one; hence, all dummy variables are retained. Dummy variables correspond to these variables, delinquency in the last 2 yrs, open accounts, public records, total accounts, and total revolving high limit are not statistically significant. Hence, all these variables are removed from the PD model. Now, the PD model only contains statistically significant sets of dummy variables.

5.2.2 PD model building using NN

The probability of the default model can be built using [2] NN as well. Figure 5 represents the Step-by-step guide to building the neural network.

The core of the NN is Artificial Neurons. Each neuron connection has an associated weight, an input function, and an output function. Weights of neurons are initialized at the beginning, but the weights are adjusted as training is progressed. The structure of NN as shown in Fig. 6 (No. of input neurons, No. of output neurons) is determined first by analyzing the data set. Here Layer 2 represents the hidden layer and

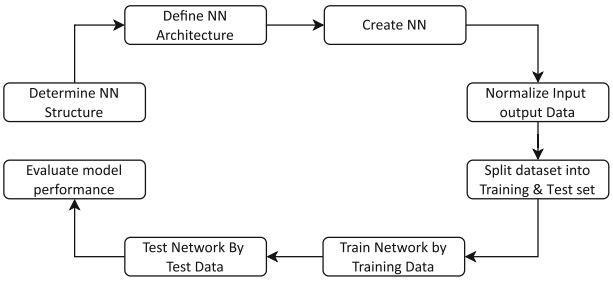

Fig. 5 Steps to develop neural network

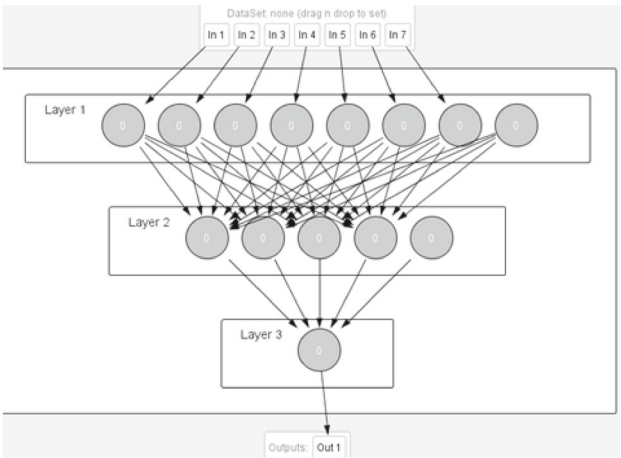

Fig. 6 Neural network structure

the number of hidden layers is 1. Multiple input signals corresponding to different input parameters like the company's financial data, and years in business are coming in NN. The input function 'weighted sum' is used to convert all those input signals into a single value type a number between 0 and 1.

The architecture of NN is designed and decided in this stage (Type of NN, No. of hidden layers, No. of neurons in each hidden layer, Learning rule, Transfer/Activation function). There are different types of NN. Here, the multilayer perceptron (MLP) type of NN is created. Multilayer perceptron NN has an input layer, an output layer, and one or more hidden layers in between. Every layer has a potentially different but fixed number of neurons in it. Here, backpropagation with momentum is chosen as a learning rule. The output function or activation function determines the value of the output signal. The sigmoid function is used as an activation function as the values in the data set are in the interval between 0 and 1. Finally, NN is created based on NN structure and architecture. Java NN framework [46] Neuroph is used to create NN.

Normalize data set Data in the input data set are in different ranges. It is not meaningful to compare data of different ranges. So, data needs to be normalized. We have split the normalized data set into the training and test data set. 70%

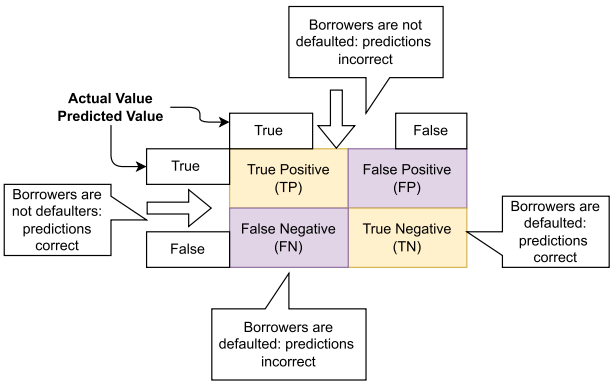

Fig. 7 Confusion matrix

data of the data set is used for training and the rest 30% data is used for testing purposes. The training data set consists of input signals assigned with the desired output.

Train neural network To train NN, learning parameters like max error, learning rate, and momentum need to be decided first. The training is complete when the Total Net Error is below the max error. The learning rate indicates the amount of change to the model during each step of the training process. The value of the learning rate is between 0 and 1.

Test neural network After training is completed, the test needs to be done against the test data set to ensure that the model is trained properly.

5.2.3 Evaluation of model performance

The performance of the model depends on the extent to which the model correctly classifies the good borrowers and bad borrowers. We can make the final classification into good or bad borrowers based on the estimated probabilities of being good or bad. We need to decide a cut-off probability. All observations with estimated probability greater than the cut-off probability are classified as good and less than or equal to the cut-off probability are classified as bad.

- Estimated Probability > Cut-off probability: Good (Non-default)
- Estimated Probability <= Cut-off probability: Bad (Default)

Confusion matrix To determine the performance of the PD model, we can determine [57] confusion matrix. The confusion matrix plays an important role to describe the performance of an ML model. The overall process is graphically represented in Fig. 7.

The confusion matrix shows the accuracy of the model. The accuracy of the model is the total number of correctly predicted observations divided by the total observations.

Table 3 Interpretation of ROC curve

Area under ROC curve	Interpretation
50% <= a < 60%	Poor
60% <= a < 70%	Fair
70% <= a < 80%	Good
80% <= a < 90%	Excellent
a >= 90%	Fair

Apart from accuracy, error rate and other statistical parameters like sensitivity, specificity, precision, and false positive rate are also computed from the confusion matrix. If the model generates a lot of false positive observations imply a lot of bad applicants would be given a loan which is not acceptable. If we take more conservative threshold, there are much false positive predictions but also much true positive. This means if lender uses this ML model for granting the application, they would reduce the number of defaults dramatically but also the number of approved applications which leads to losing business. In credit risk modeling, risk needs to be minimized but at the same time losing business is not an option. So, we can understand while measuring PD model performance, accuracy is not the only most important parameter. The rate of true positive prediction and false positive prediction is more important parameters than overall accuracy.

One common approach to show the true positive rate and false positive rate for different thresholds is the ROC curve (Receiver Operating Characteristic Curve). Every point of that curve corresponds to one threshold point that would generate a different confusion matrix. Table 3 presents a common scale for interpretation of the area under the curve.

From the ROC curve shown in Fig. 8, we can calculate false positive rates, the true positive rates, and the probability thresholds at which the respective false positive rate and the true positive rate were obtained. All of these data are useful for setting decision-making cut-offs.

Result of Probability of Default model which includes accuracy, confusion matrix and some other parameters is shown in Fig. 9. Accuracy of this model is calculated as 84.40% using NN. False positive percentage is 10.27 which indicates for 10.27% bad applicants loan is going to be granted.

PD models are turned into simplified model versions called scorecards so that they can easily interpret and understand. We need to turn the regression coefficients from our PD model into simple scores. First, we need to decide the minimum score and maximum score. Each observation falls into only one dummy category of each original independent variable. Higher coefficients represent better creditworthiness. The maximum creditworthiness assessment can get from the

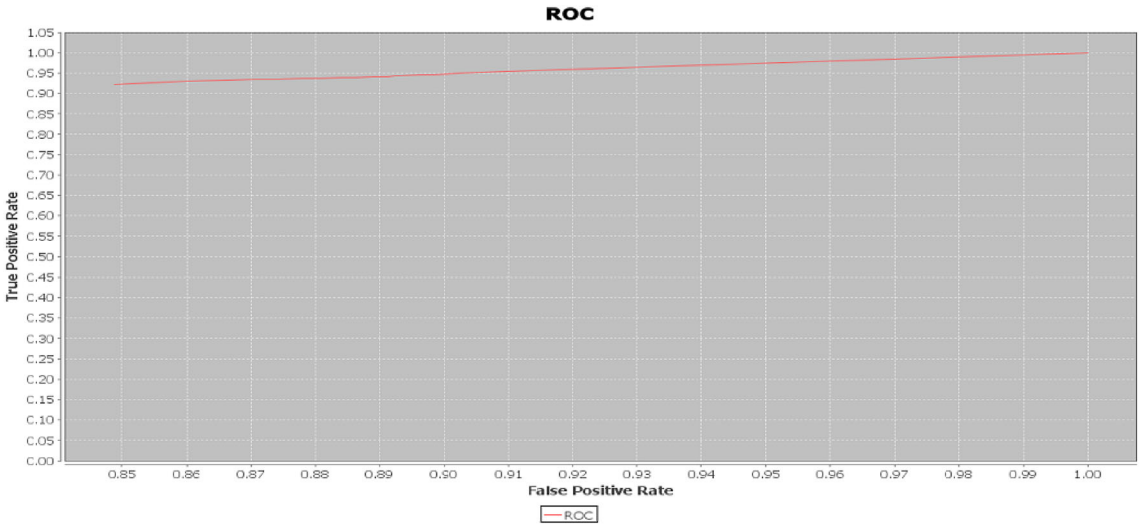

Fig. 8 ROC curve of PD model

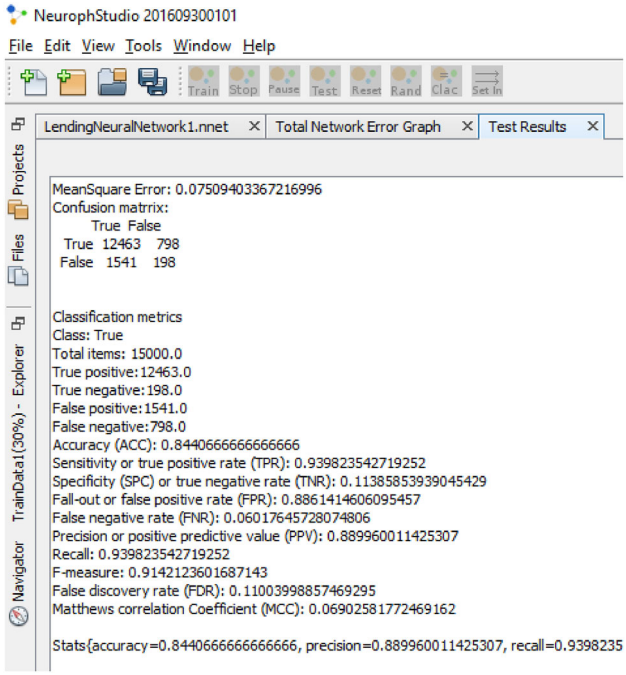

Fig. 9 PD model performance

PD model when a borrower falls into the category of original independent variables with the highest model coefficients. Similarly, the minimum creditworthiness is reached when a borrower falls into the category with the lowest model coefficients for all variables.

The scorecard is prepared by using a variable score of all independent variables. The variable score can be computed by below formula:

$$\text{variable\_score} = \text{variable\_coef} * (\text{max\_score} - \text{min\_score}) / (\text{max\_sum\_coef} - \text{min\_sum\_coef})$$

Apart from the score and estimated probabilities of default, the cut-off rate is also being used to decide whether to approve a loan application. A specific cut-off has two major implications. Once the cut-off is chosen, the total number of borrowers that will be approved and rejected is determined which impacts the quality of the loans the bank would approve. There is a trade-off between the two. If we want to approve more loans, it means the loan is approved for lower-quality borrowers. On the other side, if we want to approve the most creditworthy borrowers, then only a few loans are granted. A financial institution would not prefer this as it impacts their business. So cut-off point needs to be decided based on these two factors. If the bank wants to lend to fewer borrowers with higher credit worthiness it will set a higher cut-off point in terms of the probability of non-default. If a bank wants more business, a lower cut-off point is set in terms of the probability of non-default.

5.3 Loss given default model

Loss-given default is the share of an asset that is lost if a borrower defaults. The established approach is to model the proportion of the total exposure that can be recovered by the lender. Once a default has occurred, this proportion is called the recovery rate. The proportion that can't be recovered or loss given default can be calculated easily because it equals to (1 - recovery rate) for each exposure.

The amount received after a default is present in the column "Recoveries". We assume that in our data for defaulted borrowers the funded amount column reflects the total amount that was lost altogether at the moment the borrower defaulted. So we can calculate the recovery rate as the propor-

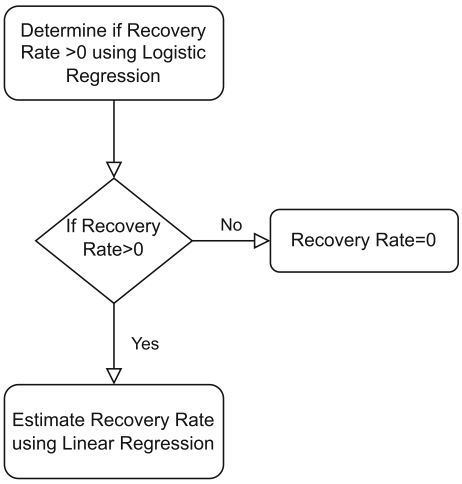

Fig. 10 Modeling steps for LGD

tion of the funded amount that has been recovered. Recovery Rate = Recoveries / Funded Amount

The recovery rate is our dependent variable for the loss given the default model. The recovery rate is restricted to intervals between 0 an 1. This specific distribution is called beta distribution. The regression model used to assess the impact of a set of independent variables on a variable with beta distribution is called Beta regression. Currently, there is no major python library that supports a stable version of beta regression. So, we need to find out an alternative approach.

From the distribution of recovery rates, we can see that about half of the observations have a recovery rate of zero while the rest of the recovery rates are greater than zero. Figure 10 represents the modeling steps of LGD. For estimating LGD, two-stage model approaches can be taken.

1. Model to determine whether the recovery rate is zero or not.
2. If the recovery rate is greater than zero, then design a model to know how much exactly it is.

The first problem is a binary question of whether the recovery rate is zero or not. So, we can use logistic regression for this model. Then we need to take only accounts where the recovery rate which is greater than 0 to estimate the recovery rate. The easiest way to do that is with linear regression. We can easily reach the final recovery rate predictions by simply multiplying the predicted values from the two models.

5.4 Exposure at default model

Exposure at default is the total value that a lender is exposed to when a borrower defaults. Therefore, it is the maximum that a lender may lose when a borrower defaults on a loan. On many occasions, the lender has decided to grant an amount of money but the lender has not dispersed the whole amount. Moreover, the borrower may be able to repay and spend what they had already repaid up to a certain limit called credit limit using revolving facilities such as credit cards. So, the borrower may only have defaulted on a proportion of the original funded amount which is going to be our dependent variable for the exposure at the default model. Most often that proportion is called the credit conversion factor. Also, the borrower may have repaid a significant amount of the debt at the time of default. Then, the exposure a default (EAD) = funded amount * credit conversion factor

The credit conversion factor is the proportion of the original amount of the loan that is still outstanding at the moment when the borrower defaulted. The total recovered principal reflects the total payments made on the principle of the loan. If a borrower defaults they would only have to repay the funded amount less the payments made on the principal as this is the money they've already repaid. So, the credit conversion factor (CCF) = (funded amount - total recovered principal)/funded amount

If all amounts have been paid, CCF would be zero. If nothing has been paid, CCF would be one. For the EAD model, credit conversion factors are more homogeneous and solely distributed. So, a linear regression can be directly applied to the model credit conversion factor. Here, for credit conversion factors, a multiple linear regression model can be used.

6 Conclusion and future work

The traditional credit assessment process is facing many challenges in handling new situations and technical demands. In this work, a solution approach, as well as a framework, is defined for the ML technique-based credit assessment system. For assessing credit risk, generally, data from loan applications, loan-related data, existing data of the borrower with the lender, and macroeconomic data are considered. In this work, an automated ETL process has been implemented so that if there is any new data in source systems that can be replicated in the DW in near real-time. In this work, three ML models namely Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EAD) are built and suggested to compute expected loss based on Basel II standards. For building models, ML algorithms, as well as NN approaches, are used. Finally, the performance of one model is also assessed.

After assessing the performance of one model, we can conclude that this new approach can be adopted for credit risk assessment to reduce risk and increase revenue. The proposed ML-based credit assessment measures the credit risk of applicants much more accurately and quickly than the traditional process. It reduces the processing time of loan applications significantly. ML-based credit assessment has

taken the advantage of use of additional data from different sources, e.g., borrower's credit card history, repayment behavior, previous financial transactions, public records of pending court cases, etc. which is not taken into consideration in the traditional process of the credit assessment.

As ML-based models consider different data dimensions, credit risk for borrowers without credit history can be determined using ML-based models. Moreover, the major benefit of this approach is that prediction is always done on real-time data as changed data is captured in near real-time through an automated ETL process.

As of now, we have taken traditional data for building ML models. As a next step, we should consider a hybrid approach in which along with the traditional data, alternative data like social media data, utility payments profile, mobile logs, GPS, and mobile usage data will also be considered to enhance the models. Also, the unstructured data which are collected by the lending institutes during day-to-day operations, for example, notes taken during interaction with customers, could be an alternate source for credit risk modeling.

Even though the prediction from the ML model reduces the processing time of lending applications significantly, the real benefit will be realized once the model uses the decision information to further reinforce its own learning. Each of these decisions will have its respective trail of information which can be used for further enhancing the prediction model to the extent that the model independently can do the prediction. The prediction system in that case does not support a human decision-making process but takes its own decisions without any human intervention or supervision.

As a future scope of the work, the parameter PD can be justified on more than two factors using other machine learning-based approaches. Although primarily our focus in the article is the application of machine learning in data integration and ETL, exploring the same in the cloud for cloud database heterogeneity can be another interesting challenge.

References

1. Abdou H, Pointon J (2011) Credit scoring, statistical techniques and evaluation criteria: a review of the literature. Intell Syst Account Financ Manag 18(2–3):59–88
2. Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H (2018) State-of-the-art in artificial neural network applications: a survey. Heliyon 4(11):e00938
3. Akkaoui EE, Zimányi E (2009) Defining ETL worfklows using BPMN and BPEL. In: Proceedings of the ACM twelfth international workshop on Data warehousing and OLAP, ACM, pp 41–48
4. Akkaoui ZE, Zimányi E, López JNM, Mondéjar JCT et al (2013) A BPMN-based design and maintenance framework for ETL processes. Int J Data Warehous Min (IJDWM) 9(3):46–72
5. Akkaoui ZE, Zimànyi E, Mazón JN, Trujillo J (2011) A model-driven framework for ETL process development. In: Proceedings of the 14th international workshop on Data Warehousing and OLAP, ACM, pp 45–52
6. Altman E, Esentato M, Sabato G (2016). Assessing Italian SME and mini-bond issuer credit worthiness
7. Altman EI (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J Financ 23(4):589–609
8. Altman EI, Sabato G (2007) Modelling credit risk for SMES: evidence from the us market. Abacus 43(3):332–357
9. Altman EI, Sabato G, Wilson N (2008) The value of non-financial information in SME risk management. J Credit Risk 6(2):1–33
10. Attigeri GV, Pai M, Pai R (2017) Credit risk assessment using machine learning algorithms. Adv Sci Lett 23(4):3649–3653
11. Bai C, Shi B, Liu F, Sarkis J (2019) Banking credit worthiness: evaluating the complex relationships. Omega 83:26–38
12. Barboza F, Kimura H, Altman E (2017) Machine learning models and bankruptcy prediction. Exp Syst Appl 83:405–417
13. Benink H, Wihlborg C (2002) The new Basel capital accord: making it effective with stronger market discipline. Eur Financ Manag 8(1):103–115
14. Bhatore S, Mohan L, Reddy YR (2020) Machine learning techniques for credit risk evaluation: a systematic literature review. J Bank Financ Technol 4(1):111–138
15. Biswas N, Chattapadhyay S, Mahapatra G, Chatterjee S, Mondal KC (2017) Sysml based conceptual ETL process modeling. In: Communications in computer and information science. International conference on computational intelligence in communications and business analytics, Springer, Singapore, pp 242–255
16. Biswas N, Chattapadhyay S, Mahapatra G, Chatterjee S, Mondal KC (2019) A new approach for conceptual ETL process modeling. Int J Ambient Comput Intell (IJACI), IGI Glo 10(1):30–45
17. Biswas N, Sarkar A, Mondal KC (2018) Empirical analysis of programmable ETL tools. In: International conference on computational intelligence, communications, and business analytics, Springer, pp 267–277
18. Biswas N, Sarkar A, Mondal KC (2020) Efficient incremental loading in ETL processing for real-time data integration. Innov Syst Softw Eng 16(1):53–61
19. Carey M (1998) Credit risk in private debt portfolios. J Financ 53(4):1363–1387
20. Castellanos M, Simitsis A, Wilkinson K, Dayal U (2009) Automating the loading of business process data warehouses. In: Proceedings of the 12th international conference on extending database technology: advances in database technology, ACM, pp 612–623
21. Chern C, Lei W, Huang K, Chen S (2021) A decision tree classifier for credit assessment problems in big data environments. Inf Syst e-Bus Manag 19(1):363–386
22. Chi G, Uddin MS, Abedin MZ, Yuan K (2019) Hybrid model for credit risk prediction: an application of neural network approaches. Int J Artif Intell Tools 28(05):1950017
23. Continuous integration-delivery-deployment in next generation data integration. https://kb.informatica.com/whitepapers/4/Documents, Accessed August 27, 2019
24. Dar AA, Anuradha N, Qadir S (2019) Estimating probabilities of default of different firms and the statistical tests. J Glob Entrep Res 9(1):1–15
25. Dastile X, Celik T, Potsane M (2020) Statistical and machine learning models in credit scoring: a systematic literature survey. Appl Soft Comput 91:106263
26. De Basilea CSB (2006) Basel ii: international convergence of capital measurement and capital standards: a revised framework–comprehensive version
27. Doumpos M, Lemonakis C, Niklis D, Zopounidis C (2019) Analytical techniques in the assessment of credit risk. EURO advanced tutorials on operational research. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-99411-6
28. Eckert J, Jakob K, Fischer M (2016) A credit portfolio framework under dependent risk parameters: probability of default, loss given

default and exposure at default. J Credit Risk 12(1). https://ssrn.com/abstract=2794935
29. Edmister RO (1972) An empirical test of financial ratio analysis for small business failure prediction. J Financ Quant Anal 7(2):1477–1493
30. El-Sappagh SHA, Hendawi AMA, Bastawissy AHE (2011) A proposed model for data warehouse ETL processes. J King Saud Univ: Comput Inf Sci 23:91–104
31. Embley DW, Campbell DM, Jiang YS, Liddle SW, Lonsdale DW, Ng YK, Smith RD (1999) Conceptual-model-based data extraction from multiple-record web pages. Data Knowl Eng 31(3):227–251
32. Ershadi M, Omidzadeh D (2018) Customer validation using hybrid logistic regression and credit scoring model: a case study. Calitatea 19(167):59–62
33. Frydman H, Altman EI, Kao D (1985) Introducing recursive partitioning for financial classification: the case of financial distress. J Financ 40(1):269–291
34. Galindo J, Tamayo P (2000) Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. Comput Econ 15(1):107–143
35. Haselmann R, Wahrenburg M (2016) Banks' internal rating models-time for a change? the" system of floors" as proposed by the basel committee. Technical report, SAFE White Paper
36. Hung K, Cheng HW, Chen S, Huang Y et al (2013) Factors that affect credit rating: an application of ordered probit models. Rom J Econ Forecast 16(4):94–108
37. Institute CF. Credit Risk. https://corporatefinanceinstitute.com/resources/knowledge/finance/credit-risk/, Accessed Mar 15, 2021
38. JR, Bernardino J (2008) Real-time data warehouse loading methodology. In Proceedings of the 2008 international symposium on Database engineering & applications, ACM, pp 49–58
39. Jones S (2017) Corporate bankruptcy prediction: a high dimensional analysis. Rev Account Studi 22(3):1366–1422
40. Lappas PZ, Yannacopoulos AN (2021) A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. Appl Soft Comput 107:107391
41. Li DX, Financial A (2000) The valuation of basket credit derivatives: a copula function approach. In: Fields institute workshop on options in financial products: approaches to valuation, Toronto, Canada
42. Maldonado S, Pérez J, Bravo C (2017) Cost-based feature selection for support vector machines: an application in credit scoring. Eur J Oper Res 261(2):656–665
43. Mondal KC, Biswas N, Saha S (2020) Role of machine learning in ETL automation. In: Proceedings of the 21st international conference on distributed computing and networking, pp 1–6
44. Muñoz L, Mazón JN, Pardillo J, Trujillo J (2008) Modelling ETL processes of data warehouses with UML activity diagrams. In: Workshops on the move to meaningful internet systems: OTM, Springer, pp 44–53
45. Muñoz L, Mazón JN, Trujillo J (2009) Automatic generation of ETL processes from conceptual models. In: Proceedings of the ACM twelfth international workshop on data warehousing and OLAP, ACM, pp 33–40
46. Neuroph. Java neural network framework. http://neuroph.sourceforge.net/. Accessed Mar 15, 2021
47. Oliveira B, Belo O (2012) BPMN patterns for ETL conceptual modelling and validation. In: Foundations of intelligent systems, Springer, pp 445–454
48. Radhakrishna V, SravanKiran V, Ravikiran K (2012) Automating etl process with scripting technology. In: Nirma university international conference on engineering (NUiCONE), IEEE, pp 1–4
49. Rizvi NU, Kashiramka S, Singh S (2018) Basel I to Basel III: Impact of credit risk and interest rate risk of banks in India. J Emerg Mark Financ 17(1–suppl):S83–S111
50. Shen F, Zhao X, Li Z, Li K, Meng Z (2019) A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation. Phys A: Stat Mech Appl 526:121073
51. Shumway T (2001) Forecasting bankruptcy more accurately: a simple hazard model. J Bus 74(1):101–124
52. Skoutas D, Simitsis A (2006) Designing ETL processes using semantic web technologies. In:Proceedings ACM 9th international workshop on data warehousing and OLAP (DOLAP 2006). Arlington, Virginia, USA, pp 67–74
53. Skoutas D, Simitsis A (2007) Ontology-based conceptual design of ETL processes for both structured and semi-structured data. Int J Semant Web Inf Syst (IJSWIS) 3(4):1–24
54. Song J, Bao Y, Shi J (2010) A triggering and scheduling approach for ETL in a real-time data warehouse. In: Computer and information technology (CIT), 2010 IEEE 10th international conference on,IEEE, pp 91–98
55. Soui M, Gasmi I, Smiti S, Ghédira K (2019) Rule-based credit risk assessment model using multi-objective evolutionary algorithms. Exp Syst Appl 126:144–157
56. Suresh S, Gautam JP, Pancha G, DeRose FJ, Sankaran M (2001) Method and architecture for automated optimization of ETL throughput in data warehousing applications. US Patent 6,208,990
57. Susmaga R (2004) Confusion matrix visualization. In: Intelligent information processing and web mining, Springer, pp 107–116
58. Trujillo J, Mora SL (2003) A UML based approach for modeling ETL processes in data warehouses. LNCS, Springer Verlag 2813(2003):307–320
59. Uddin MS, Chi G, Al Janabi M, Habib T (2022) Leveraging random forest in micro-enterprises credit risk modelling for accuracy and interpretability. Int J Financ Econ 27(3):3713–3729
60. Van Roy P (2005) Credit ratings and the standardised approach to credit risk in Basel ii. ECB Working Paper
61. Varotto S (2008) An assessment of the internal rating based approach in Basel ii. Journal of Risk Model validation
62. Vassiliadis P, Simitsis A (2008) Near real time ETL. Springer annals of information systems, 3(978-0-387-87430-2). Special issue on New Trends in Data Warehousing and Data Analysis
63. Vassiliadis P, Simitsis A, Skiadopoulos S (2002) Conceptual modeling for ETL processes. Proc DOLAP, pp 14–21
64. Zhou H, Yang D, Xu Y (2011) An ETL strategy for real-time data warehouse. In: Practical applications of intelligent systems, Springer, pp 329–336