ISSN 2278-3091
Volume 10, No.3, May - June 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse1281032021.pdf
https://doi.org/10.30534/ijatcse/2021/1291032021

Assessment of Default Risk Factors in the Disbursement of Home Loans

Zafar Nasir[a], Zeeshan Ahmed, Chaman Lal[b,c]
[a]Faculty of computing and Information Technology, Indus University Karachi Pakistan
[b]Faculty of Engineering, Science and Technology, Indus University Karachi Pakistan
[c]Mohammad Ali Jinnah University Karachi Pakistan
zafarnasir@indus.edu.pk, zshnaz@yahoo.com, chaman.lal@indus.edu.pk

ABSTRACT

Considerable amount of time and effort is required to assess and evaluate the financial credit risk inherent in the specific request for the award of home loans, especially in the private sector. It has been a challenging scenario for the financial institutions to ascertain the financial strength of the prospective customer to pay back the loan amount in a stipulated time frame. This estimate is critical to ensure the financial viability and profitability of the enterprise entrusted with the obligation to disperse the financial credit. A binary decision system that is capable to analyze in a few seconds whether a loan applicant is financially viable / suitable for issuance of the loan amount he has requested for, can revolutionize the loan disbursement mechanism. Insufficient or non-verifiable credit history is the major hurdle in accurate prediction of bad debts and recovery rates of the loans committed by the financial institutions. For the purpose of research within the scope of this work, data-sets have been utilized, with data points gathered together by a certain 'Home Credit', that are stored in files of CSV (Comma Separated Values), that houses a diverse set of information on the basis pertains to lender's willingness to grant the loan and the other part relates to borrower's ability to repay the loan.

Many methods do exist, but are not quite perfect, to challenge the rate of rejection and acceptance criteria for a credit lender's decisions for the better. For this research's take, the focus is shifted on the datasets provided, and maintained, by the financial loan provider, Home Credit Group. Understanding the role of repaying a loan as the ebb and flow of growing business model, Machine Learning algorithms of time frames, and nature of the loans. Naturally, noise is a recurring factor, as the data sets are generally found to be imbalanced, noisy, and heterogeneous.

To dissemble the complication at large, Machine Learning Algo rhythms, which lean to using pre-processing techniques, are availed to explore, analyze and determine the crucial factors that play together in the projection of a risk. In addition, the manipulation of the K-Nearest Neighbors (KNN), and a neural network with ensemble learning have worked out fairly well in this case by incorporating specific, important individual features. Each feature is incorporated as a future-weight directly proportional to the entropy of the feature. Initial comparison of the results with the state-of-the-art, tried and tested results, have given the impression that the proposed technique scores higher than already present and in-use models of classification.

Key words: Dataset, Loan default, Credit history, Default Risk, Data-frame.

1. INTRODUCTION

Moving away from today's list of financial calamities, to simply just focus on the loan aspect of it, it is seen that the overall context of consumer credit can be perceived in two parts: The first part have offered new ways for features in data, as those not already seen, or shared, in traditional datasets, all aimed at driving the development of modern statistical methodologies. Existing modern datasets feature massive sample sizes with a great order of dimensionality, far more than what was dealt with in traditional procedures. It is important to understand the role of large sample sizes, and the relation of heterogeneity before moving forward.

Firstly, large sample sizes mean, small sub-populations and weak communities that exist within the entire community. This could be, on one hand, hard to achieve as developing models of intrinsic heterogeneity of dataset requires far more complex, computational heavy statistical methods. Secondly, as the incoming data requires collection of a user's annual income, credit history, bank balance, other specific loans etc. It is indeed possible to arrive at the idea of noisy data, spurious correlation, and imbalances between past and present values, all the while the main target being to find if a certain loan might be a risk or not. All these forms the basis of a major problems of Deep Learning and Machine Learning, as data collected might change, vary, might not be in the required format, may contain outliers, that quickly disorganize a dataset, while the research is aimed at making the best of the historical data in our efforts to evaluate the main problem; does a person qualify for a loan or not?

This problem thus assumes the role of a classification problem with provided labels and certain input layers, thus also

enabling us to apply the principles of Supervision for better and thorough estimation of the solution. Defined below are two terms, 'classification', and 'supervision', that will make the assumptions and techniques applied more plausible.

- Supervised: In the training dataset if labels are given, it means we have to train our model and predict the labels from the features and evaluate the result with the help of some suitable evaluation metric.

- Classification: The labels that are given in the dataset, 0 and 1. 0 means loan is successfully repaid on time and 1 means difficulty to repay his loan.

In this study, We are using 3 different machine-learning methods (Logistic Regression, Decision Tree and KNN), 2 Ensemble Learning (Random Forest and Light Gradient Boosting), deep neural network with different activation functions and different number of hidden layers, modified version of (KNN and decision tree) and ensemble learning with deep neural network pre-processed by KNN with some values of K, find the k nearest neighbors to predict the default risk based on client's pre-processed data, and compare their prediction in term of accuracy, Precision, Recall, F1 score and Confusion Matrix. By & large the overall performance of the techniques in general is in contrast of their potential to predict the default risk using multiple indicators (accuracy, precision, recall and F-score). The outcomes received indicate that the classification ability of neural network with ensemble learning and Light-GBM is most appropriate to different machine-learning techniques which include neural networks. It is additionally located that the overall performance of neural-network models relies upon on the preference of activation function and the range of middle layers. This chapter introduced the basic problem of loan prediction. What are the chances they will pay their loan or not? The remaining sections arranged as follows:

Section 2: Analyses the literature and related work in classification of loan repayment abilities, with particular focus on techniques that have been chosen for the classification of loan applications in this work.

Section 3: Mainly describes the methodology that has been chosen for work in this study, for classification of loan application from Home Credit Group. It describes the overall classifiers framework as well as implementation details. This also includes, implementation of feature selection, one hot encoding, feature engineering, feature ordinal, imputer, transform the training and testing data using Min-Max Scaler and various machine learning algorithms for classification, the deep learning model, neural networks with its layer, their loss and activation function, optimizer and dropouts, regularization and validation techniques used in this work along with overview of the dataset used and implementation aspects. Summarizes the dataset which is provided by Home Credit that contains multiple CSV files and these are collected from different time frames and varies greatly in nature. The challenges of dataset are imbalanced, noisy, heterogeneous, merging the dataset and curse of conditionality. Also defines the proposed approach. Section 5: Reviews the key experimental result of this study. Section 6: Conclusion.

2. RELATED WORK

In terms of research done, a large body of literature can be found regarding monetary organizations. This would-be consistent with, and not limited to, financial, fiscal, and budgetary managers, lenders, governments, and financial market players which provide insight into the fundamental basis in what constitutes the development of efficient models, and effectively, estimating an approximation into the likelihood of loan defaults. Speaking scholastically, information that deals with capital market can be helpful in determining the models that predict bankruptcy. As a guide, take [1], which a groundbreaking and noteworthy paper, puts into application the use of multivariate statistical techniques, predominantly into the department of discriminant analysis, to provide classification on solid, secured and prolific companies suing financial statement data. To improve on this idea, researchers have tried, including [18], who developed pioneering ways to analyze dataset and apply logistic regression for default estimation. Contrasting with [1], which is determined to classify between beneficial and non-beneficial payers, the work of is said to describe the likelihood of the potential borrower.[18]

Owing to the modern conventions and convenience, the con-lenience, easiness and over-all general usability of performing Discriminant Analysis intensely on tried and tested data sets, along with the hand in hand inclusion of Logistic Regression, have led diverse, divergent, and various studies that have come after having sought to execute, carry out, tests with more or less the same kind of results [14], [12] and [4]. Thought, it should be noted that [2] raised an objection to that the much favored models based on [1] and [18], presenting their reason, citing that they had become inaccurate with the growing needs, incapable of producing the desired variations, and remarked suggestions that talk in detail over the need for enhancements in the modeling of welshing risk.

Those belonging to the fields of Pedagogy, along with many in the field of academia, counting practitioners, researchers, and the modern banking system are trying to find ways to open, expand, and bring to new horizons the tools of Artificial Intelligence and Machine Learning to calculate credit risk in the midst of leaping, rapid and fast-moving advances in field of general computer technology. Credit risk analysis has been found to be quite similar to pattern-recognition problems, thus allowing many already existing propositions to be used in the classification of the creditworthiness of counterparties [15] [19]. The aforementioned algorithms have improved traditional, and cultural models based on straightforward and painless Multivariate Statistical Techniques. As an example, one can take the case of discriminant analysis with Logistic Regression. A new breath has been ushered with other, newer, methods being put into development with the intent of modifying using optimization. These new investigations have

been part of a monumental effort and research, offering brand new substitutes or alternative than already existing ones for highlighting the criterial preferences of credit risk analysis. e, we also highlight machine learning methods. SVM, with different kernels [6], taken as an example with the regard to their inspection of data, generate formulas and mechanisms, which are in fact quite parallel, matching and similar in comparison, to discriminant analysis, but owing to the reasons of them not being specific and subject to a varying degree of assumptions, this has led them to being less restrictive.

Additional Machine Learning methods do exist that exhibit wide applicability and the ability to handle such kind of workloads. To push further, oracular and revelatory models have nonetheless also been the subject of proposals for improvement, including the basic and the simple default models. To name these, it would be boosting, bagging, and random forest models. Artificial Neural Networks (ANN) have been the source of generating much intrigue and interest in many contexts as well. Such incorporation has led, at least, for these machine learning algorithms to have so far seemed encouraging and reassuring. For a specimen, take [17], which tapped into the Australian, German, and Japanese financial datasets, and then to have come to conclusions to be indicative of the facts that Machine Learning Techniques, such as the already mentioned ensemble methods, have more or less been the lead for a better way for classification [24], than to simply have invested in some standalone methods.

While it appears that a multitude of studies have analyzed the solvency of the corporate world through the use of modern computational techniques. But even in the face of such profound data, it was realized that the outcomes and ramifications did not assist nor did they avail much in reaching out the helping hand to provide classification for the identification of the best methods. This was due to the fact that the model performance depended much, almost entirely, on the precise, discrete, or characteristics that would be kept in particular mind during the totality of the process involved in the classification problem, while also on the mention of data structure [10]. Furthermore, [23] used ensemble methods, to name a few - bagging, boosting, stacking – stacked and joined with the base learners, which are, again – Logistic

Regression, Decision Trees, ANN, and SVM - to have come to the conclusions and find out that bagging actually managed to outperform boosting when talked in regards with each and every one credit databases that were analyzed by the academics.

A range of differing, dissimilar and heterogenous monitors, surveys and examines have dealt with the fiery dialogues and exchanges, adding more the heated discussions and talks that are concerned over the strengths, the weaknesses of Machine Learning in many, many opposed and contrasting disciplines, such as [22] and [8] in medicine; [16] and [20] in chemistry;

[3] in education; and [5], [13], [21] and [11] in the financial world. However, our study adds more constructive insight to provide and give back some contribution to this debate.

To add firstly, it is indeed true that our study focuses on the certain, specific, comparisons of old-fashioned, accustomed, even conventional statistical methods and Machine Learning Techniques. These comparisons do deal with the workflow, the performance, and the reasons behind why new, modern, and revolutionary changes are a must. It should also not be forgotten that these comparisons provide intuition, prescience and apprehension for the prediction of fiscal discrepancies, such as corporate bankruptcy. Although some papers have brought to light and have put into motion studies the issue, their cons and pros, the advantages and disadvantage of credit defaults and their interactions with the whole world and business of Machine Learning [7] [9];state-of-the-art, fresh, and modernistic probes and technological journeys, exploring different models, contexts and datasets, are very relevant, because the results that prevailed through the models, were mostly regarding the superiority of models, and have led us to believe that they are still inconclusive. The controversy, and contention and the main quarrel over the most suited models for the job, will the ones responsible for the selection of the numbers that will eventually be in charge of the prediction and with being the forbearers of the news of failure, and that will probably continue, starting at the minimum from the short range, and all the way up to medium range and any other terms that may as such lie in between. This is far in contrast, when spoken about in context of rivalled models and contemporary techniques which are repeatedly and continually being conspicuous, with a slight detail to note in, which would be to be particularly aware for the study of corporate bankruptcy. Other such events, such as those of failure events, are the spine and the main supporting thesis to many of the droves and masses of otherwise variables that pop up in such circumstances. In such occurrences and factors, taken as an example, with the promotion, growth, and preferment that technological field has become well known with. These units vested with the responsibility for the tasks of data scraping will do good to the field through the surveillance, consideration and monitoring of new, automatic and dynamic variables that for sure are very relevant inputs to machine learning models, and which may eventually lead to many spin-off results, with each result having a different direction of its own.

Secondly, with the extension available for a wide variety of assortment of models, methods, processes and techniques of applicability provided to, by, and for practitioners, it can also be considered as positive impression, immersion, and as a token of gratitude towards the target of the study. Through the extensive use of raw data and by putting into deep consideration of standardized, normalized and assimilated computer knobs and switches in place already for the machine learning techniques, all our models can be easily replicated, not only by researchers in the field of academics, but by market practitioners as well to a varying degree. A much-needed discussion also entails that these models can be put under the use and implementation of such real-world situations, as like those already present in the world-like problems, already grounded in reality, with situations to mark with the purpose of inscribing in stone and paper the ideas much need for the development and contrivance. Talking

archetypically, the case that speaks about the problems and calculations, the predictions and skills required thereof from investors, could perhaps be better understood and that would provide much needed in-depth analyses relating to strategic credit decisions.

Keeping in view the factors discussed earlier, it would be prudent to discuss in depth the case of lender institutions, which to many also seem to be engrained with the foresight to help them improve their credit risk controls, arising from the root firmly established on consistent outputs and derivatives. Moreover, the yield procured and extracted from Machine Learning Models manifest their existence all through the sequence of valuable data points. Our work scrutinized the performance of different classification expertise by considering differing applications of practical problems of default prediction, under administration of Machine Learning Algorithms. We have, for this purpose, utilized the dataset of Home Credit Group, and use some evaluation metrics like accuracy, precision, recall, f1 score and confusion matrix to compare the performance of algorithms.

3. METHODOLOGY

3.1 The Dataset

Data used to explain the relationship between credit loan payment and some financial and demographic variables were obtained through consumers (individual) credit records of the Home Credit. This is an international non-bank financial institution. Dataset is publicly available on Kaggle. They are openly challenging to help them to unlock the full work- able of their data. Doing so will make sure that consumers successful repayment is not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their customers to be successful. The data is provided by Home Credit, this service is provided to less privileged population, and predicting whether or not customer will repay their loan or having issues. Because it's a critical business need and Home credit group is hosting this competition on Kaggle with the prize money of $70,000 and they want to use some machine learning and feature engineering expertise, provides the best results. This dataset contains the 8 tables. Each table is given in .CSV format. The

Data-frame application train has 307511 rows and 122 columns, the data-frame application test has 48744 rows and 121 columns, the data-frame bureau has 1716428 rows and 17 columns, the data-frame bureau balance has 27299925 rows and 3 columns, the data-frame credit card balance has 3840312 rows and 23 columns, the data-frame installments payments has 13605401 rows and 8 columns, the data-frame previous application has 1670214 rows and 37 columns, and POS CASH balance has 10001358 rows and 8 columns. The dataset is completely imbalanced, which contains 91.9% takes the values 0 for paid samples and 8.07% takes the values 1 for unpaid samples. Dataset has a lot of missing values, that can drastically impact the machine learning model's quality. The loan amount ranges from 45,000 to 4,050,000. They grant the 3 types of loan: Revolving loan, Consumer loan and Cash loan. The purpose of the loan is to buy the car or a home. Most of the loan applicant's educational background is secondary and higher education; higher applicant's marital status is married; mostly applicant's occupations are Laborer, Sales staff, Drivers, Core staff and Manager; the higher applicants have their own house or apartment; mostly people are unaccompanied; higher number of applicants worked as Business Entity and Self Employed.

3.2 Tables Description

Data Architecture Diagram Fig. 1. shows the interrelation-ships between the data files provided. Following are brief explanation of each

3.2.1 Application train test.csv:

The main table contains training and the testing data with information about each loan application at Home Credit Every loan has its own row and is identified by the feature SK ID CURR. The training data-frame has TARGET columns, indicate 0: the loan is successfully repaid or 1: the loan was not repaid.

3.2.2 Bureau.csv:

Bureau data-frame has the previous loan applications. Which contains that taken from the other financial institution. One row in application data has multiple rows or previous loan applications in the bureau. It is identified by SK ID BUREAU.

3.2.3 Bureau balance.csv:

In Bureau Balance monthly data is given of those loans that are taken from other financial institutions. Each row is one month of a previous application. It means this table contains multiple rows of the previous one loan application. It has a one-to-many-relationship.

3.2.4 Previous application.csv:

Previous applications contain those loans which are taken from the same organization (Home credit default risk). One row application data in application data frame have multiple rows in previous application and it is identified by the column SK ID PREV.

3.2.5 POS CASH balance.csv:

Monthly data about previous applications in terms of point of sale or cash loan that took from the same institute. Every row is one month of a previous cash or sale loan. And single previous loans have multiple rows.

3.2.6 Credit card balance.csv:

Monthly data about the previous credit card clients. Every row is the one month of previous credit card balance.

3.2.7 Installments payments.csv:

Previous loans payment history at Home credit group. There is one row of every payment or missed payment.

3.2.8 Home Credit columns description.csv:

In this file description of every features that is given in data-frames.

This diagram shows overall structure of data-frames and shows how these are related.

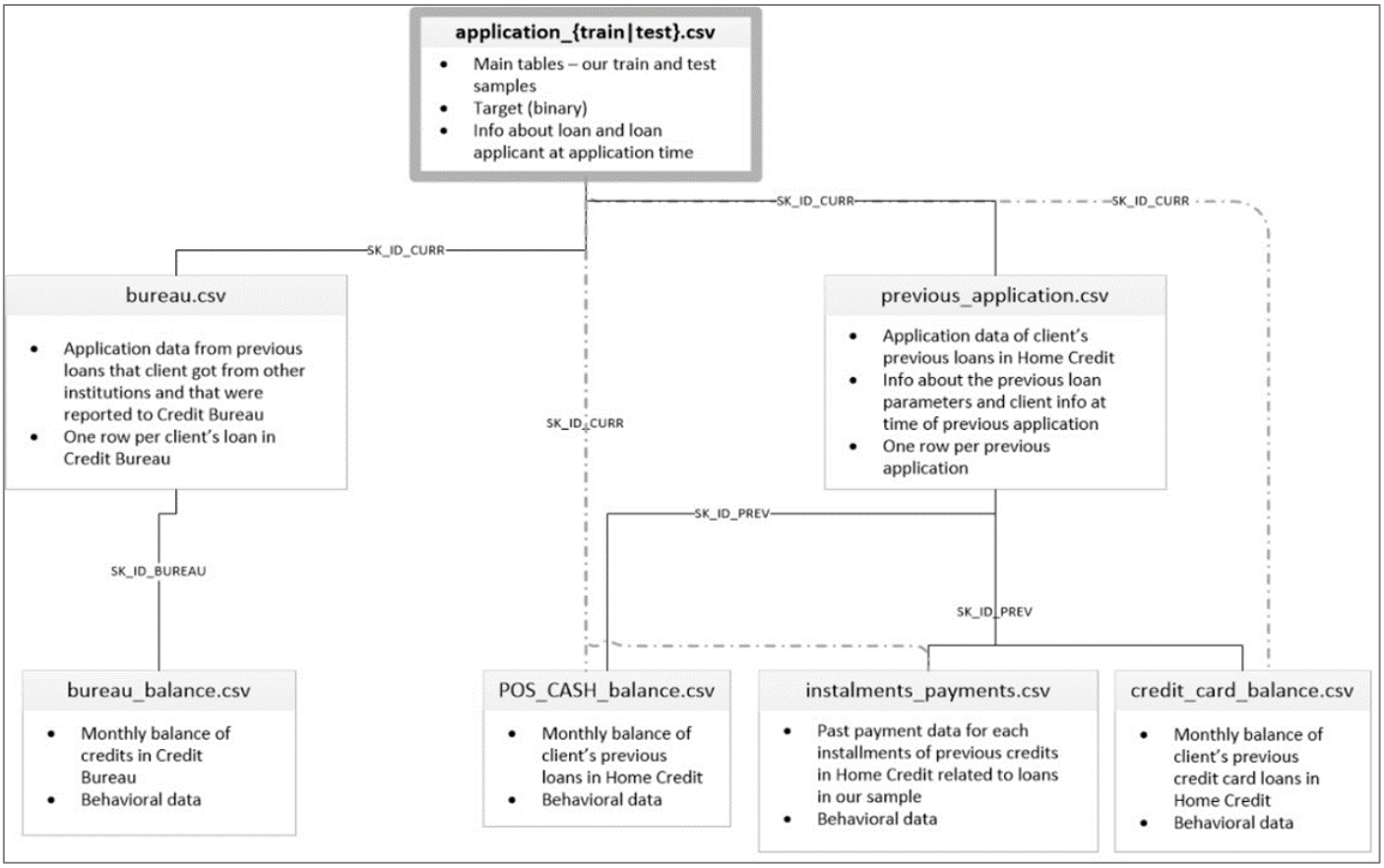

Figure. 1: Dataset Description.

4. EXPERIMENTAL STUDIES

In this work, A dataset has been developed in which almost 7 CSV files from many different sources that may change and alter in midst of the data collection process in two classes; repaid and defaulter. Firstly, a substantial amount of time examining their raw data. This study involves, producing and visualizing salient features, remove redundant features, handle outliers and missing values in the data for identification of any anomalies or outliers that exist in the data. Using feature extraction techniques (e.g., PCA) for normalization on the features.

After that some baselines model (e.g., Decision tree, Logistic Regression, Random Forest, KNN and Light GBM) is implemented on filtered features. Accuracy, precision, recall, F1 score and confusion matrix is used as an evaluation measure.

4.1 Proposed Approach:

The detailed procedure of the financial credit risk evaluation is generally described in Fig. 2. The data underneath examination comes from the financial variables of companies. After pre-processing (filtering, filling missing value, transformation, normalization etc.), feature extraction and feature engineering tools are used for dimension reduction using PCA or other techniques that are given in the previous research. Later, statistical, intelligent machine, ensemble and deep learning algorithms are applied. Deep domain knowledge is integrated with modeling algorithm to support the investigation and comprehension of the data. The proof consists of the value preferences of the users, the class label of investigated data or observation as properly as some large, privileged information. Initially Decision Tree, Linear Regression and Random Forest is adopted as our base model, for learning. Later advanced state-of-the-art classifiers is investigated. Precision, Recall, Accuracy, F1 score and confusion Matrix is used as an evaluation matrix.

4.2 Pre-processing and Cleaning:

Our dataset is highly imbalanced, which has a lot of missing values and redundant features are the major challenges. we then pre-processed our dataset in 2 different ways with little variation. The dataset is given in 8 different .csv files. The data-frame application contains the loan and loan applicants. The data-frame bureau contains the application data from other loans that the client took from other financial institutions and were reported to the credit bureau. Previous applications of data-frame contain information about previous loans that he took from the Home Credit by the same client SK ID CURR is connecting the data-frames application train—test with bureau, previous application and also with data-frames POS CASH balance, installments payment and credit card balance. SK ID PREV connects data-frame previous application with POS

CASH balance, installments payment and credit card balance. SK ID BUREAU connects data-frame bureau with data-frame bureau balance. The dataset contains data-frame application train and application test contain the loan and loan applicants. The data frame bureau contains the application data from other loans that the client took from other financial institutions and were reported to the credit bureau. The data-frame previous applications contain information about previous loans that he took from the Home Credit by the same client, previous loans information and client information at the time of the loan (there is a line in the data-frame per previous loan application). SK ID CURR is connecting the data-frames application train—test with bureau, previous application and with data- frames POS CASH balance, installments payment and credit card balance. SK ID PREV connects the data-frame previous application with POS CASH balance, installments payment and credit card balance. SK ID BUREAU connects data-frame bureau with data-frame bureau balance. The dataset contains the 3 different types of values such as object, int64 and float64. The newest applicant may apply for loan to other organization, the previous applicant data given in bureau.

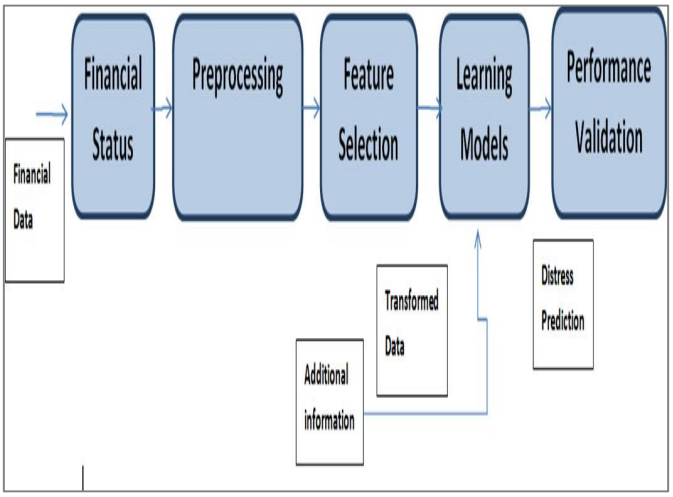

Figure. 2: Pipeline of financial credit risk assessment.

1) First Technique:

We have 7 different CSV files including train and test separate files. We have to implement the preprocess steps on train and test because machine learning and deep learning wants to same dimension of train and test data. Training data is used to train the model and test data is totally unseen data for model. Performance is evaluated in terms of evaluation metric. Data comes from several different sources, and in this sub step we focus on integrating these different sources. Data varies in size, type and structure in csv files and enriching an observation from one table with information from another table using left join. The overall framework of data preprocessing is shown in Figure 3. After that, the ID columns are removing for example SK ID CURR, SK ID BUREAU and SK ID PREV. Now we have 377 at- tributes. Calculate the correlation of each feature and set the threshold value is 0.9, those attributes which full fill this condition, it means they are highly correlated and have the similar information, these are redundant features remove them. Now we have 291 attributes. Find the missing values of each feature in terms of percentage, those features which have missing values greater than 75% remove them, the rest of the features are 246. Calculate the average and mode of each column and fill the entire data-set with average or mode. This is first technique of pre-processing of our data. The pre-process framework is given in Figure 3.

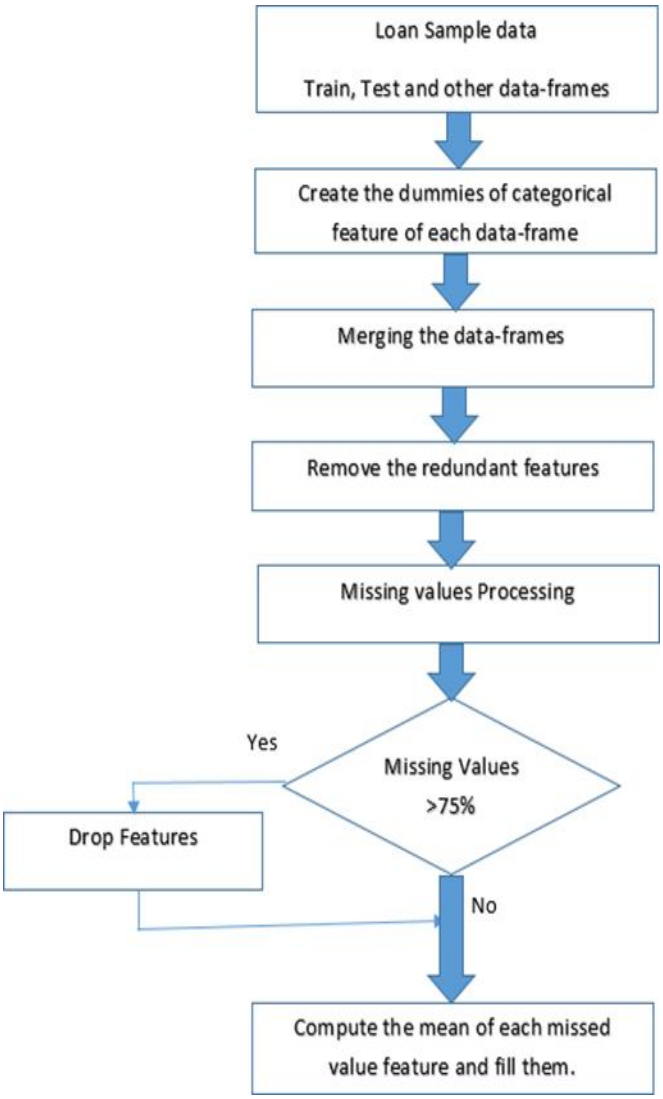

Figure. 3: First Data pre-processing framework.

2) Second Technique:

The 2nd technique is a modified version of the 1st technique. We are also using the same dataset as in 1st technique. In this technique we also create the dummies of categorical features of each data-frame using one hot encoding. Which means convert the one column which has n number of categories, we will use n number of columns into n vectors. These vectors are 1s and 0s to represent whether the category value occurs or not. Implement feature engineering which improves the prediction and performance of models. We also calculate the

correlation between feature to feature and feature with target. We also drop the categorical and redundant features. Merging the data-frames. Sometimes we may find attributes which have missing values and outliers in the dataset. We need to handle these anomalies and fix this problem. Obviously, we removed the entire line of instance or attribute but sometimes unknowingly removing crucial information. But in this case, we find the missing values in terms of percentage. Those attributes which have missing values greater than 75%. Remove them, rest of the features take a mean of all values of each column and replace the value on missing data position in the same column. And implement the Min Max Scaler technique, transform the training and testing data. The pre-process framework is given in Figure 4.

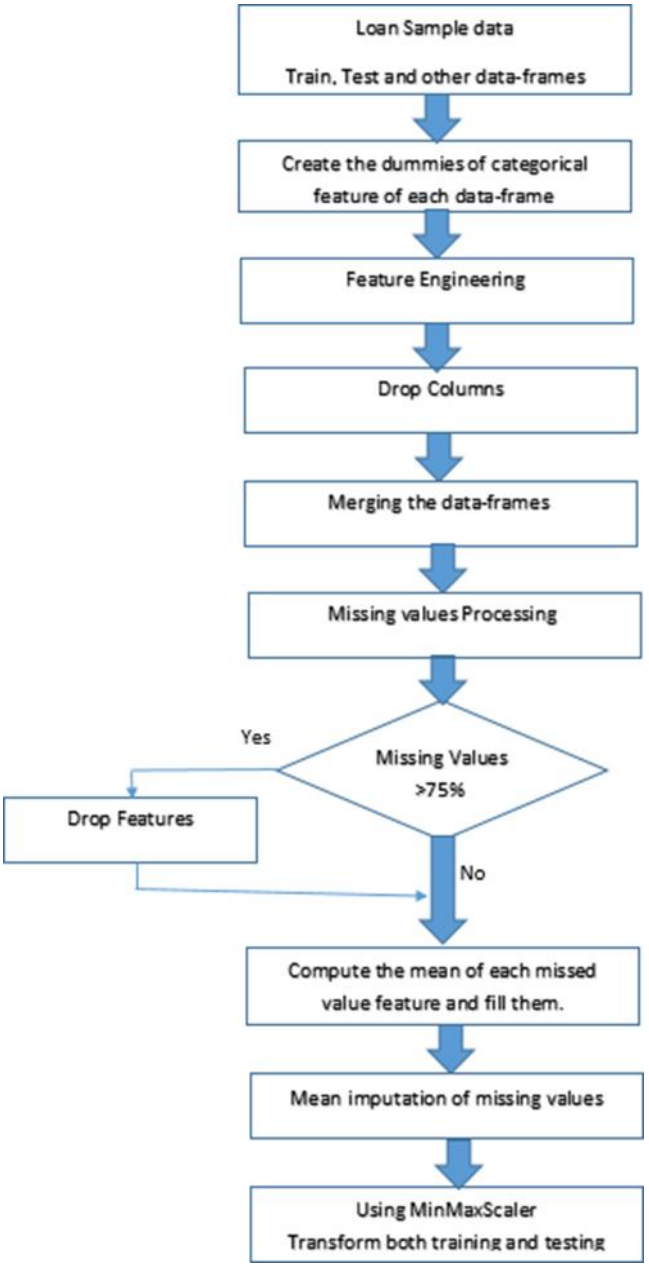

Figure. 4: Second Data pre-processing framework.

4.3 Classification

1) Logistic Regression:

We are using all the features after encoding the categorical variables and merging the dataset. In this section, we pre-processed the data by first technique filling in the missing values (imputation) and normalizing or transforming the range of the features by Min Max Scaler. We are using Logistic Regression from Scikit-Learn for our first model. The only change is to the model from the default model settings is to lower the regularization parameter, C, which controls the amount of overfitting (a lower value should decrease overfitting) and we used the C= 1e15. This provides slightly better results than the default Logistic Regression, but even than it sets a low bar for any future model Here we use the familiar Scikit-Learn modeling syntax: We firstly create the model, then we train the model using. fit and then we make the predictions on the testing data using. predictive probability. (Remember that we want probabilities and not a 0 or 1). The logistic regression baseline should score around (private: 0.70221 and public: 0.69982) when submitted on Kaggle. But we also split the dataset into 70% and 30%. 70% for training and 30% for testing and make the prediction on entire test data and the accuracy is 91.96%, Recall is 99.98%, Precision is 91.97% and f1 score is 95.80%.

Table: 1 CONFUSION MATRIX OF LOGISTIC REGRESSION

	Predicted Positive	Predicted Negative
Condition Positive	84829	12
Condition Negative	7403	10

2) Decision Tree:

Decision Tree is a supervised learning algorithm which splits the attribute for each node of a tree and builds a binary tree. A decision tree algorithm is built from a top-down structure, which starts from the root node and splits the data into subsets that belong to the same class. Entropy is a measure of disorder-ness or homogeneity or how messy our data? A homogeneous subset means entropy is 0 it means sample distribution belongs to a single class. The information gain at each level is calculated by using the entropy of the parent node and the weighted sum of entropy of its child nodes. Info gain value decides which node is split further. After training model builds, the tree has a set of rules decided. For unseen data or test data, follow these rules and predict the class label. The Decision Tree baseline model score around (private: 0.68627 and public: 0.69018) when submitted on Kaggle. The result of the ROC curve is 0.63. But we also split the dataset into 70% and 30%. 70% for training and 30% for testing and make the prediction on entire test data and the accuracy is 92.01%, Recall is 99.48%, Precision is 92.01% and f1 score is 95.59%.

Table: 2 CONFUSION MATRIX OF DECISION TREE

	Predicted Positive	Predicted Negative
Condition Positive	84402	439
Condition Negative	7321	92

3) Random Forest:

Dataset is pre-processed by first tech- unique. Random forest is a tree-based ensemble learning algorithm that generates randomly multiple Decision Trees by using random subset of the data, which is just like forest. Pick the sample from test data, follow the rules of this algorithm, and achieve class labels by individual decision trees which are the part of forest and predict the class label with majority vote. Random forest is an ensemble learning algorithm which helps to overcome the over-fitting problem. It significantly increases accuracy of the model.

Table: 3 CONFUSION MATRIX OF RANDOM FOREST

	Predicted Positive	Predicted Negative
Condition Positive	84830	11
Condition Negative	7410	3

This model should score around (private: 0.72703 and public: 0.72890) when submitted. The result of the ROC curve is 0.72. But we also split the dataset into 70% and 30%. 70% for training and 30% for testing and make the prediction on entire test data and the accuracy is 91.955%, Recall is 99.98%, Precision is 91.96% and f1 score is 95.80%.

4) K-Nearest Neighbor:

K Nearest Neighbor method can create both classification and regression models as well. KNN is a pattern recognition algorithm. In this method we calculate the distance from the test point to all training points in feature space. Find the K closest points and see the labels. Those K nearest samples belong to majority class, it means we have to classify the test sample into majority class. Because it has a majority of votes. This is an instance-based algorithm. This algorithm is a very simple classification algorithm when we have no idea about distribution of data, this is very slow. But the algorithm tricky part is the value of K. In this work we are using k values from 1 to 9, because we have no idea about the best value of K. So, the superior performance of KNN is K=8. In this technique every sample is classified as its surrounding samples. The model should score around (private: 0.72010 and public: 0.72563) when submitted. The result of the ROC curve is 0.73. But we also split the master dataset into 70% and 30%. 70% for training and 30% for testing and make the prediction on entire test data and the accuracy is 91.89%. And we also implement the same model on 30,000 samples the accuracy is 91.75% on k=4.In another experiment randomly choose the 30,000 samples on a random basis and split the sample dataset into 70% and 30%. 70% for training and 30% for testing and make the prediction on entire test data and the accuracy is 89.19% on k=8%, ROC is 0.57, Recall is 99.95%, Precision is 91.41% and f1 score is 95.48%.

Table: 4 CONFUSION MATRIX OF NEAREST NEIGHBOR on K=8 with 30,000 Samples

	Predicted Positive	Predicted Negative
Condition Positive	8223	4
Condition Negative	772	1

5) Light Gradient Boosting Machine

Now we use a real machine learning model: the gradient boosting machine using the Light-GBM library! The Gradient Boosting Machine is currently the leading model for learning on structured datasets. The volume of data is increasing day by day and it is very challenging to manage, maintain and give the faster results with accuracy for traditional data science algorithms. Light-GBM is prefixed as" Light" because of its excessive speed. Light-GBM can handle a rich dataset and use less resources (less amount of memory) to run. Another reason for LGBM popularity is, it focuses on best results as well as speed and uses less memory. Light-GBM supports GPU learning and data scientists mostly use this algorithm in data science problems. The model should score around (private: 0.77673 and public: 0.77684) when submitted. The result of the ROC curve is 0.7815. But we also split the dataset into 70% and 30%. 70% for training and 30% for testing and make the prediction on entire test data and the accuracy is 91.96%, Recall is 99.83%, Precision is 92.09% and f1 score is 95.80%. When we use the data, which is pre- process by second technique the model score on leader board is private: 0.79851 and public: 0.80028.

Table: 5 CONFUSION MATRIX OF LIGHT GRADIENT BOOSTING MODEL

	Predicted Positive	Predicted Negative
Condition Positive	84703	138
Condition Negative	7273	140

6) Proposed Methods using KNN and Decision Tree:

In this proposed approach, firstly we calculate the entropy of pre-processed data.

En = -Pn * log (Pn, 2) - Qn * log (Qn, 2) n=1,2

Merging the entropies of each attribute.

Ex = (weight of En * En) + (weight of En * En)

Secondly, we compute the Euclidean distance between each test point with every train data point using entropy.

distance = math.sqrt (math.sum (math.square (Ex)* math.square (x test - X train[i, :])))

Sort the distances.

distances = sorted (distances)

Make the list of K's nearest neighbor's target.

for i in range(K):
index = distances[i][1]
targets.append(y train.iloc[index])
Return most common target.
Counter(targets). most common (1)[0][0]

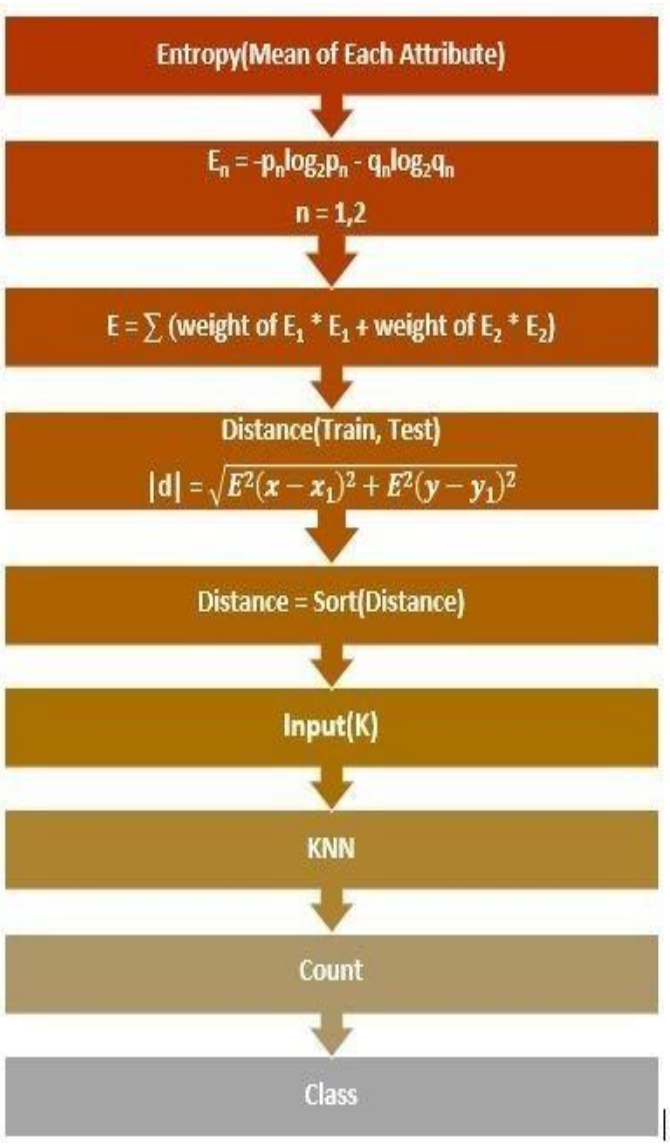

Figure. 5: Modified Version of K-NN with.

Decision Tree This approach is implemented on 30,000 samples, we also split the dataset into 70% and 30%. 70% for training and 30% for testing and make the prediction on entire test data and the accuracy is 91.91%, Recall is 99.75%, Precision is 92.11% and f1 score is 95.77%.

Table: 6 CONFUSION MATRIX OF KNN+DT on k=8 with 30,000 samples

	Predicted Positive	Predicted Negative
Condition Positive	8270	20
Condition Negative	708	2

7) Neural Network:

A neural network is a mathematical model that is impressed by the knowledge processing of the human brain. It is an interconnection of many neurons. The uniqueness of this model is based on the fact that it incorporates past experiences and take a decision in a given amount of time. Neural network's basic structure is 3 layers: the first layer is input layer; one or more number of hidden layers and one is output layer. We -have to feed the raw data or input data into the input layer, hidden layers processing on input data and show the output on the output layer. Every neuron of the previous layer is fully connected with next upcoming layer neurons. In the hidden layer every unit has its own weight which is any number during iterations they update till the network achieve the appropriate output based on input data. Finally, we show the output on the output layer that is generated by the hidden layer using input data and activation function.

Table: 7 CONFUSION MATRIX OF Neural Network with random under-sampling by first pre-processed Technique

	Predicted Positive	Predicted Negative
Condition Positive	17814	7011
Condition Negative	6887	17938

8) Neural Network Implementation:

Our deep neural network architecture consists of one input layer with input shape, 400 neurons and ReLU Activation Function Six hidden layer with 1024,2048,1024,2048,712 and 356. These neurons. activation function ReLU, Batch normalization and dropout 0.6 have been used after each hidden layer. One output layer with activation function Sigmoid. In this network we use the Adam as an optimizer with parameters learning rate is 0.001, beta 1 is 0.9 and beta 2=0.999. Loss function is binary cross entropy and batch size is 512.

After pre-processed by first technique we implement some sampling techniques. For example, we implement Random under-sampling after that we implement deep neural network, score around (private: 0.73603 and public: 0.74021) when submitted on Kaggle. But we also split the dataset into 70%, 10% and 20%. 70% for training and 10% for validation and 20% for prediction the accuracy is 72.0%, Recall is 7175.25%, Precision is 72.11% and f1 score is 71.92%.

After implementing SMOTE and deep neural network, score around (private: 0.67778 and public: 0.66740) when submitted on Kaggle. But i also split the dataset into 70%, 10% and 20%. 70% for training and 10% for validation and 20% for prediction the accuracy is 91.01%, Recall is 82.44%, Precision is 99.49% and f1 score is 90.16%.

```
model = Sequential([
  Dense(400, activation='relu', input_shape=(350,)),

  Dense(1024, activation='relu'),
  BatchNormalization(),
  Dropout(0.6),

  Dense(2048, activation='relu'),
  BatchNormalization(),
  Dropout(0.6),

  Dense(2048, activation='relu'),
  BatchNormalization(),
  Dropout(0.6),

  Dense(1024, activation='relu'),
  BatchNormalization(),
  Dropout(0.6),

  Dense(712, activation='relu'),
  BatchNormalization(),
  Dropout(0.6),

  Dense(356, activation='relu'),
  BatchNormalization(),
  Dropout(0.6),

  Dense(1, activation='sigmoid')])

  model.compile(Adam(lr=0.001, beta_1=0.9, beta_2=0.999, epsilon=1e-08,
decay=0.0),loss="binary_crossentropy",metrics=['accuracy'])

  es=EarlyStopping(monitor='val_loss', mode='min', verbose=1)

  model.fit(X_train1, y_train1,validation_data=(X_val,y_val),batch_size=512, epochs=4000,
verbose=1, callbacks=[es])
```

Figure. 6: Deep Neural Network Code.

Table: 8 CONFUSION MATRIX OF Neural Network with SMOTE sampling by first pre-processed Technique

	Predicted Positive	Predicted Negative
Condition Positive	233048	281507
Condition Negative	1179	49638

In the last after implementing Near-Miss and deep neural network, score around (private: 0.59235 and public: 0.58422) when submitted on Kaggle. But we also split the dataset into 70%, 10% and 20%. 70% for training and 10% for validation and 20% for prediction the accuracy is 59.26%, Recall is 57.82%, Precision is 96.55% and f1 score is 72.32%.

Table: 9 CONFUSION MATRIX OF Neural Network with Near Miss sampling by first pre-processed Technique

	Predicted Positive	Predicted Negative
Condition Positive	166921	121765
Condition Negative	5954	18871

After pre-processed by the second technique we implement some sampling techniques. For example, implementing Random under sampling, after that we implement a deep neural network, score around (private: 0.76960 and public: 0.77695) when submitted on Kaggle. But we also split the dataset into 70%, 10% and 20%. 70% for training and 10% for validation and 20% for prediction the accuracy is 71.88%, Recall is 69.50%, Precision is 73.49% and f1 score is 71.43%.

Table: 10 CONFUSION MATRIX OF Neural Network with Random Under-sampling by second pre-processed Technique

	Predicted Positive	Predicted Negative
Condition Positive	17669	7755
Condition Negative	6373	18451

After implementing SMOTE and deep neural network, score around (private: 0.67611 and public: 0.70094) when submitted on Kaggle. But we also split the dataset into 70%, 10% and 20%. 70% for training and 10% for validation and 20% for pre- diction the accuracy is 87.69%, Recall is 75.97%, Precision is 99.22% and f1 score is 86.05%. In last, after implementing.

Table: 11 CONFUSION MATRIX OF Neural Network with SMOTE sampling by second pre-processed Technique

	Predicted Positive	Predicted Negative
Condition Positive	214761	67915
Condition Negative	1672	281010

Near Miss and deep neural network, score around (private: 0.50366 and public: 0.5010) when submitted on Kaggle. But we also split the dataset into 70%, 10% and 20%. 70% for training and 10% for validation and 20% for prediction the accuracy is 23.93%, Recall is 18.09%, Precision is 95.57% and f1 score is 30.42%.

Table: 12 CONFUSION MATRIX OF Neural Network with Near Miss sampling by second pre-processed Technique

	Predicted Positive	Predicted Negative
Condition Positive	51156	231550
Condition Negative	2369	22456

4.4 Neural Network with Ensemble Learning:

In this pro- posed approach, firstly we have to pick the pre-processed data from the second technique, after that we pick the data sample one by one and calculate the distance from all

data points. We arrange them in descending order. We enter the value of K. And pick the K nearest neighbors with labels and arrange them row wise using NumPy concatenate. Using the above technique, we modified our train and test dataset. We have implemented the stratified K Fold cross validation with 25 splits. 1 part is used for validation and 24 part for training. On each fold the test data labels are predicted. After 25- fold cross validation. We have 25 predictions; we take the majority vote using mode or mean. And finally classify the data points given below.

1) In this proposed approach, firstly we implement the random under-sampling on whole dataset.
2) Calculate the distances of every individual point to all points.
3) Sort the distances in ascending order.
4) Using stratified K-fold with 25 splits with 6 layers neural network, 1 for validation and 24 for training. After each fold predict the test data and save in individual array.
5) After completing 25- fold, take the majority vote, finally assign the label to the test point.

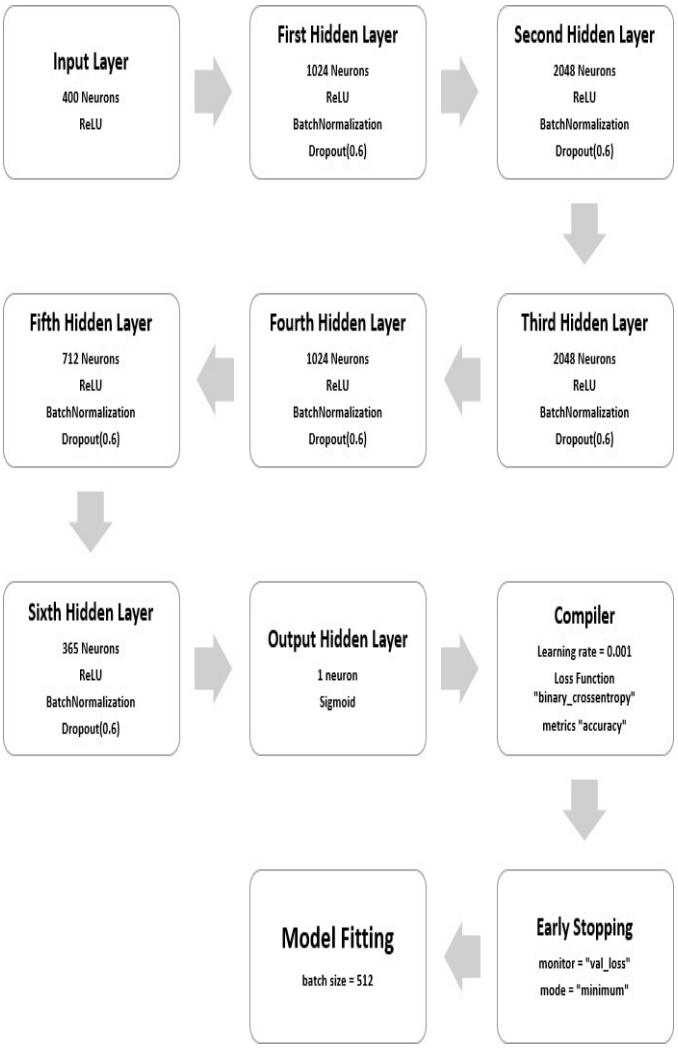

Figure. 7: Deep Neural Network Architecture.

Above modified version of the Neural network represents the set of policies used for predicting. The Neural Network pre-processed by KNN score on k=3, majority vote by mean around (private: 0.56851 and public: 0.56934), majority vote by mode (private: 0.56049 and public: 0.56526) and on k=5, majority vote by mean around (private: 0.57503 and public: 0.56867), majority vote by mode (private: 0.50000 and public: 0.50000). But when submitted on Kaggle. We also split the dataset into 70% and 30%. 70% for training 10% for validation and 20% for testing and make the prediction on entire test data and the accuracy is 99.99%, Recall is 99.98%, Precision is 100% and f1 score is 99.99%.

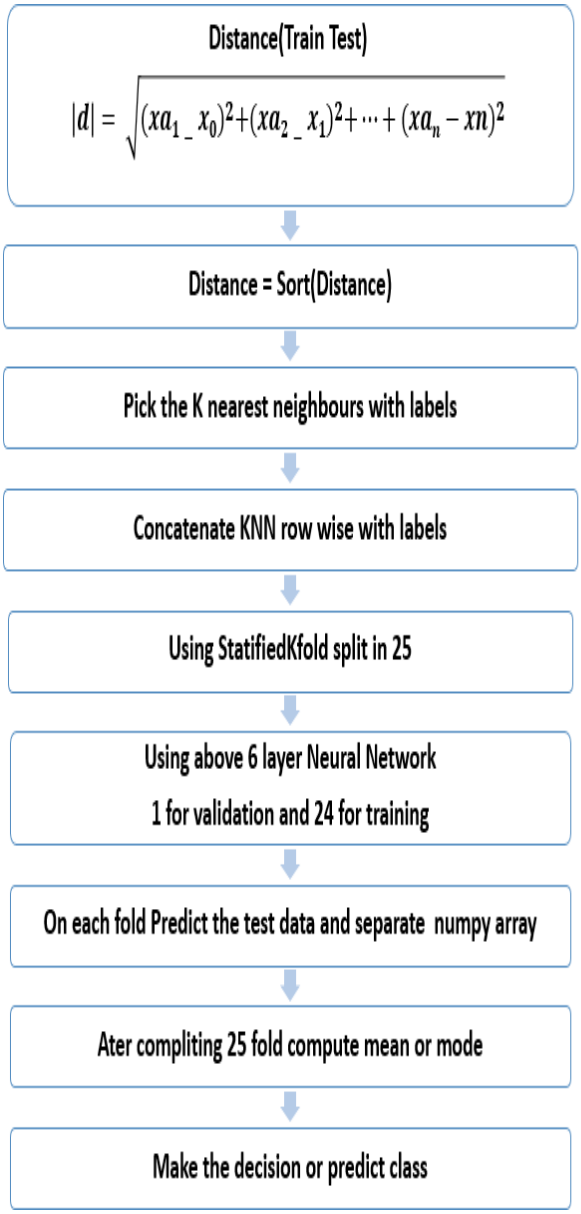

Figure. 8: Deep Neural Network with Ensemble Learning Architecture.

Table: 13 CONFUSION MATRIX OF Neural Network with ensemble Learning pre-processed by KNN on K=5 split into 70 and 30.

	Predicted Positive	Predicted Negative
Condition Positive	51156	231550
Condition Negative	2369	22456

If we split our dataset into 60 and 40. 60% for training and 40% for testing. Prediction on the entire test data, accuracy is 99.98%, Recall is 99.96%, Precision is 100% and f1 score is 99.98%.

Table: 14 CONFUSION MATRIX OF Neural Network with ensemble Learning pre-processed by KNN on K=5 split into 60 and 40.

	Predicted Positive	Predicted Negative
Condition Positive	9931	0
Condition Negative	3	9925

Table: 15 CONFUSION MATRIX FOR Neural Network with ensemble Learning pre-processed by KNN on K=5 split into 50 and 50.

	Predicted Positive	Predicted Negative
Condition Positive	12413	0
Condition Negative	1	12410

5. EXPERIMENTAL RESULTS

Before training and testing, we split our master training data into 70:30. 70% for training and 30% for testing. We evaluate the performance of our test classification model. After that finally we predict the actual test data and submit the predicted result on Kaggle. The Machine learning classifiers are used like, Decision Tree, Logistic Regression, K nearest Neighbor, Modified Version Of KNN. Ensemble Learning Classifiers are Random Forest and Light GBM, Deep Neural Network with 6 hidden layers. Pre-processed data using k=5 nearest neighbors, Deep Neural Network with ensemble learning. 10-fold cross validation is observed for accuracy, precision, recall and f1 score for classifier. Optimized results were obtained by fine-tuning the different parameters. In 10-fold cross validation, the whole dataset is divided into 10 equal sets randomly. Then one set is considered for testing and the rest of nine sets for training. Neural Network with ensemble learning using 25 stratified K Fold cross validation. We Stratified version of KNN and pre-processed by KNN using Euclidean distance. In our work superior performance classifiers are Light-GBM and deep Neural Network with ensemble learning in terms of accuracy, precision, recall, f1 score and Kaggle results. The main objective was to give a comparative analysis of features and algorithms according to their performance.

Table: 16 Classification Accuracy, precision, Recall and F1 score obtained from different Models on 70 by 30 data distribution.

Model	Preprocess	Model Name	Accuracy	Precision	Recall	F1 Score
Model 1	1	Logistic Regression	91.96%	91.97%	99.98%	95.80%
Model 2	1	Decision Tree	91.58%	92.01%	99.48%	95.59%
Model 3	1	Random Forest	91.95%	91.96%	99.98%	95.80%
Model 4	1	K- Nearest Neighbor(k=8) with 30,000 samples	89.19%	91.41%	99.95%	95.48%
Model 5	1	Light Gradient Boosting Algorithm	91.96%	92.09%	99.83%	95.80%
Model 6	1	KNN+DT (k=8) with 30,000 samples	91.91%	92.11%	99.75%	95.77%
Model 7	1	Neural Network with random under-sampling	72.0%	72.11%	71.75%	71.92%
Model 8	1	Neural Network with SMOTE sampling	91.01%%	99.49%	82.44%	90.16%
Model 9	1	Neural Network with Near Miss sampling	59.26%	96.55%	57.82%	72.32%
Model 10	2	Neural Network with random under-sampling	71.88%	73.49%	69.50%	71.43%
Model 11	2	Neural Network with SMOTE sampling	87.69%%	99.22%	75.97%	86.05%
Model 12	2	Neural Network with Near Miss sampling	23.93%	95.57%	18.09%	30.42%
Model 13	2	Ensemble Learning pre-processed by KNN when k=5	99.98%	100%	99.96%	99.98%

6. CONCLUSION

The Default Risk concerned in Home Credit Group has been minimized, by means of identifying the good set of features all through data pre-processing and using the parameter tuning of algorithm like Decision Tree, Random Forest, KNN, LGBM and KNN+DT, Neural Network and Neural Network with ensemble pre-processed by KNN to achieve high accuracy, precision, recall, accuracy and F1 score. Although Neural network with ensemble learning pre- processed by KNN using random under-sampling has the highest accuracy 99.99%, precision 100%, recall 99.99% and f1 score 99.99% which is higher than the others on whole master dataset and Logistic Regression has 0.75 ROC-AUC. Taking into consideration random 30,000 samples of master dataset pre- processed by first technique, our proposed approach has accuracy around 92.477%. We have also shown that deep learning features demonstration can also have significant impact on the overall accuracy in predicting the outcome.

REFERENCES

[1] Edward I Altman. **"Financial ratios, discriminant analysis and the prediction of corporate bankruptcy"**. The journal of finance, 23(4):589– 609, 1968.

[2] Joy Begley, Jin Ming, and Susan Watts. **"Bankruptcy classification errors in the 1980s: An empirical analysis of altman's and ohlson's models"**. Review of accounting Studies, 1(4):267–284, 1996.

[3] Jason Bernard, Ting-Wen Chang, Elvira Popescu, and Sabine Graf. **"Learning style identifier: Improving the precision of learning style identification through computational intelligence algorithms"**. Expert Systems with Applications, 75:94–108, 2017.

[4] Jing Chen, Lor a´n Chollete, and Rina Ray. **"Financial distress and idiosyncratic volatility: An empirical investigation"**. Journal of Financial Markets, 13(2):249–267, 2010.

[5] Laura Cleofas-S a´nchez, Vicente Garc´ıa, **"AI Marque´s, and Jose´ Salvador Sa´nchez. Financial distress prediction usin the hybrid associative memory with translation"**. Applied Soft Computing, 44:144–152, 2016.

[6] Corinna Cortes and Vladimir Vapnik. **"Support-vector networks"**. Machine learning, 20(3):273–297, 1995.

[7] Paulius Danenas and Gintautas Garsva. **"Selection of support vector machines-based classifiers for credit risk domain".** Expert Systems with Applications, 42(6):3194–3204, 2015.

[8] Fortunato S de Menezes, Gilberto R Liska, Marcelo A Cirillo, and Ma´rio JF Vivanco. **"Data classification with binary response through the boosting algorithm and logistic regression"**. Expert Systems with Applications, 69:62–73, 2017.

[9] Philippe du Jardin. **"A two-stage classification technique for bankruptcy prediction"**. European Journal of Operational Research, 254(1):236–252, 2016.

[10] Edgar A Du e´n˜ez-Guzma´n and Michael D Vose. **"No free lunch and benchmarks".** Evolutionary Computation, 21(2):293–312, 2013.

[11] Eduardo A Gerlein, Martin McGinnity, Ammar Belatreche, and Sonya Coleman. **"Evaluating machine learning classification for financial trading: An empirical approach"**. Expert Systems with Applications, 54:193–207, 2016.

[12] John M Griffin and Michael L Lemmon. **"Book-to-market equity, distress risk, and stock returns"**. The Journal of Finance, 57(5):2317–2336, 2002.

[13] Junyoung Heo and Jin Yong Yang. **"Adaboost based bankruptcy forecast- ing of korean construction companies"**. Applied soft computing, 24:494– 499, 2014.

[14] Stephen A Hillegeist, Elizabeth K Keating, Donald P Cram, and Kyle G Lundstedt. **"Assessing the probability of bankruptcy"**. Review of account- ing studies, 9(1):5–34, 2004.

[15] Jochen Kruppa, Alexandra Schwarz, Gerhard Arminger, and Andreas Ziegler. **"Consumer credit risk: Individual probability estimates using machine learning"**. Expert Systems with Applications, 40(13):5125–5131, 2013.

[16] Dipak Laha, Ye Ren, and Ponnuthurai N Suganthan. **"Modeling of steelmaking process with effective machine learning techniques"**. Expert systems with applications, 42(10):4687–4696, 2015.

[17] Loris Nanni and Alessandra Lumini. **"An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring"**. Expert systems with applications, 36(2):3028–3033, 2009.

[18] James A Ohlson. **"Financial ratios and the probabilistic prediction of bankruptcy".** Journal of accounting research, pages 109–131, 1980.

[19] Rudrajeet Pal, Karel Kupka, Arun P Aneja, and Jiri Militky. **"Business health characterization: a hybrid regression and support vector machine analysis"**. Expert Systems with Applications, 49:48–59, 2016.

[20] Stephen Ross, Anthony Bossis, Jeffrey Guss, Gabrielle Agin-Liebes, Tara Malone, Barry Cohen, Sarah E Mennenga, Alexander Belser, Krystallia Kalliontzi, James Babb, et al. **"Rapid and sustained symptom reduction following psilocybin treatment for anxiety and depression in patients with life-threatening cancer: a randomized controlled trial"**. Journal of psychopharmacology, 30(12):1165–1180, 2016.

[21] Jae Yong Shin, Sung-Choon Kang, Jeong-Hoon Hyun, and Bum-Joon Kim. **"Determinants and performance effects of executive pay multiples: Evidence from korea"**. ILR Review, 68(1):53–78, 2015.

[22] Abdulhamit Subasi and M Ismail Gursoy. **"Eeg signal classification using pca, ica, lda and support vector machines"**. Expert systems with applications, 37(12):8659–8666, 2010.

[23] Gang Wang, Jinxing Hao, Jian Ma, and Hongbing Jiang. **"A comparative assessment of ensemble learning for credit scoring"**. *Expert systems with applications*, 38(1):223–230, 2011.

[24] Tkatek, Said, et al. "**Artificial intelligence for improving the optimization of NP-hard problems: a review**." International Journal of Advanced Trends Computer Science and Applications 9.5 (2020).