Financial Advisory Assistant Platform

1Meera Ambre , 2Sayali Naik, 3Rahul Nair, 4Unmesh Sawant , 5Mrs.Aarti Bakshi

1,2,3,4Student,

1,2,3,4Department of Electronics and Telecommunication,

1,2,3,4K.C.College of Engineering and Management Studies and Research, Kopri, Thane, India

5Assistant Professor,

5Department of Electronics and Telecommunication,

5K.C.College of Engineering and Management Studies and Research, Kopri, Thane, India

1ambre.meera@gmail.com , 2naiksayali31@gmail.com , 3rahulnair2798@gmail.com 4unmeshsawant55@gmail.com , 5artigauri@yahoo.com

*Abstract*— In today’s world the banking sector is facing a tremendous increase and issue regarding the non-performing loans/assets from the their customers which results in jeopardizing effect on the growth of the institute in banking world. In world, where technology is advancing daily, is have been easy for companies to store the huge data of the customers which represent their behavior. With the help of the data collected from a leading credit provided to unbanked population, we have done loan default prediction. In this paper we will see how the loan default prediction is done using four different machine learning algorithms named Naive Bayes’ Theorem, Deep Learning using four and five layers, Logistic regression and Gradient boosting. The algorithm model evaluation is done using confusion matrix, Receiving Operating Characteristic charts, Cumulative charts, etc. The evaluation also has important metrics as accuracy, sensitivity, precision, etc. After comparing the performances of the algorithm, we save the model to the disk using Pythons pickle model and make use of it for predicting the new data. This paper provides basis to find the risky customers from the bunch of applicants.

*Keywords*—Loan default, Credit, Algorithms, Evaluation.

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

# **Introduction**

Credit lending or loan lending plays significant role in the banking sector. But increase in the non-performing loans have made banking sector face huge loses and also has impact on the economy of the country or world. In research it has been found that individuals around the globe take loan to fulfill their economic constraints in order to meet their personals goals which may be regarding their education, medical, travelling, business, etc. As the loan lending and receiving is quite beneficial for both the lender and receiver, the amount of loan applicants is increasing day by day and with this increases the risk of bad credit. While giving loan many aspects of the costumer like his age, occupation, assets, etc. are taken into consideration before lending him/her a loan.

Through number of prediction models, we can help the financial institutions get to understand the economical behavior of the consumer from the various data provided. Numerous studies have been done to find such loan defaulters using various methods in machine learning. Manjeet Kumar in [7] used Neural Network to predict loan default prediction and concluded that Neural Network performed better than traditional networks. Li Yung in [8] compared three models; Random Forest, Logistic Regression, Support Vector Machine and concluded that Random Forest performed better task. Abhishek Tiwari in [11] in his default prediction used Logistic Regression, Random Forest, K- Nearest Neighbors Classifier and Classification and regression Tress and compared their evaluation to find that Random Forest and K-Nearest Neighbor Classifier have the highest accuracy. In the competition held by Kaggle[2] , Jay Borkar compared six different algorithm to find out that Gradient Boosting is best followed by Deep Learning and Random Forest . He also used many evaluation techniques to compare the model.

 In this paper we have used Naive Bayes’ Theorem, Logistic Regression, Deep Learning using different layers and Gradient Boosting algorithm. The evaluation of the models is done using Confusion Matrix, Cumulative Gain Chart, Lift Chart, Kolmogorov-Smirnov Chart, and Receiver Operating Characteristic Chart of each model. The further paper is organized as follows- Section II consists of the data interpretation and processing, Section III consists of the information of the Algorithm used, Section IV consists of Evaluation and Comparison of the model output based on different matrices. In Section V we will discuss how to display the model on the web and in Section VI discuss about the result and findings.

# **Data sources and processing**

The data for this paper was provided by Home Credit a company which provides financial inclusion for unbanked population which was available on Kaggle for a competition. Home Credit wanted to unlock the full potiential of their data to check what sort of learning models can be developed to find the defaulters. The data sources are of seven different types:

 **1) Application train and test data**- It has the main training and testing data which consist of each loan application of Home

Credit. It training data comes with 0 and 1 representing the loan was repaid and loan was not repaid respectively.

2) **Bureau**- It contains the client’s previous credit data from previous financial institutions.

3) **Bureau balance**-It has the monthly data of the previous credits in bureau.

4) **Cash balance**- It has the monthly data of sale or cash loans of Home Credit.

5) **Credit Card Balance**- It has the monthly data of the credit card holders of Home Credit.

6) **Previous Application**- It has previous applicant’s data who had applied for loan and have loan with Home Credit.

7) **Installments** – It has history of every loan repayment and missed payment.

#  **Data exploration and examining the missing values**

It is the process to find trends, similarities, patterns, relationships, traits, etc. within the data. It is done to gain as much as information from the data. The data exploration helps in the modeling choices and decide which features to use.

In the data available we need to look for the missing values in each column. We will fill these missing values with a median value which we calculate from the available data.

#  **Categorical variables and most valuable plot**

Categorical variables are those who appear frequently in the data, but all machine learning models cannot deal with these variables so we need to encode them before we use them in the model. We use Label encoding and One-hot encoding. Label encoding assigns a unique category variable with an integer. Here no new columns are created. It is done for the categorical variables with only two categories. One-hot encoding creates a new column for each unique category. It assigns 1in the column for its corresponding category and 0 in all other new categories.

When we plot the histogram to show the distribution of the numerical variable and categorical variable. This plots are called Most Valuable Plot. This helps us in the visualization of the predictability power of the predictor. They show us the variables affecting each other or the target.

# **Machine learning algorithms**

#  **Logistic regression.**

 It is the most widely used algorithm for classification purpose. The strength of the logistic regression is that it has simple understanding and easy implementation. Logistic regression is better than the linear regression as it predicts the probability of an outcome which have only two values and forms a logistic curve that is limited between 0 and1 whereas the later predicts the values outside the acceptable range of values such as negatives. As shown in the **Fig.1** the logistic regression *b0* moves the curve towards right and left and the slope *b1* shows the stiffness of the curve. The logistic regression equation can be written in terms of an odd ratio, therefore Equation 1 is

*b0*+*b1x*)

Taking natural log on both sides. The coefficient *b1* is the amount log odds change with a change in one unit of x*,* therefore Equation 2 is

*b0* +*b1x*

 We get following Logistic regression equation that can handle number of numerical and categorical variables. Therefore Equation 3 is

*p=*



**Figure 1-Linear And Logistic Regression.**

# **Naïve Bayes Theorem**

 Naïve Bayes being easy to build and having iterative parameter estimation makes it useful for the large datasets. It often surprisingly do well and is used widely. In [], Naïve Bayes classifier the value of predictor (x) on a given class (c) is independent of the other predictors values. This is called class conditional independence. Equation 4 is,

P(c |x)=P(x1|c)(x2 |c)…… P(xn | c)P(c)

Where,

 P(c |x) –It is the posterior probability of class given predictor.

 P(c) – It is prior probability of class.

 P(x |c) – It is the probability of predictor given class.

 P(x) – It is the prior probability of class.

The class with the highest posterior probability is the outcome of prediction.

# **Gradient boosting.**

Gradient boosting could be a machine learning technique for regression and classification issues, that produces a prediction model within the style of an ensemble of weak prediction models, usually call trees. It builds the model during a stage-wise fashion like different boosting strategies do, and it generalizes them by permitting improvement of an arbitrary differentiable loss operate. The idea of gradient boosting originated within the observation by Leo Breiman that boosting may be understood as AN improvement formula on an acceptable price operate. Specific regression gradient boosting algorithms were afterwards developed by Eusebius Hieronymus H. Friedman, at the same time with the additional general practical gradient boosting perspective of Llew Mason, eating apple Baxter, Peter Bartlett and Marcus Frean. The latter two papers introduced the read of boosting algorithms as repetitive practical gradient descent algorithms. That is, algorithms that optimize a price operate over operate house by iteratively selecting a operate (weak hypothesis) that points within the negative gradient direction. This practical gradient read of boosting has semiconductor diode to the event of boosting algorithms in several areas of machine learning and statistics on the far side regression and classification.

# **Deep Learning**

 A deep neural network (DNN) is a man-made neural network (ANN) with multiple layers between the input and output layers. The DNN finds the right mathematical manipulation to show the input into the output, whether or not or not it's a linear relationship or a non-linear relationship. The network moves through the layers hard the chance of every output. DNNs will model advanced non-linear relationships. DNN architectures generate integrative models wherever the article is expressed as a stratified composition of primitives. The additional layers modify composition of options from lower layers, doubtless modeling advanced knowledge with fewer units than an equally activity shallow network. DNNs area unit usually feed forward networks within which knowledge flows from the input layer to the output layer while not iteration back. At first, the DNN creates a map of virtual neurons and assigns random numerical values, or "weights", to connections between them. The weights and inputs area unit increased and come an output between zero and one. If the network didn't accurately acknowledge a selected pattern, AN algorithmic rule would change the weights. That manner the algorithmic rule will ensure parameters additional influential, till it determines the right mathematical manipulation to totally method the information.

**Deep Learning with four layers.**

It has two hidden layer and each hidden layer has 80 neurons. Dropout layer is added between the 2 hidden layer and between the second hidden layers. The input layer has 242 dimension for the data, the hidden 2 layers acts as activation function with 80 neurons each. As we are considering binary classification, the sigmoid activation function is used

**Deep Learning with five layers.**

It has three hidden layers, first and second hidden layer has 80 neurons, while the third has 40 neurons. . Dropout layer is added between the hidden layers and between the third hidden layer and the output layer. The input layer has 242 dimension for the data, the hidden 3layers acts as activation function with rectifier. As we are considering binary classification, the sigmoid activation function is used at the output layer.

# **IV Model evaluation techniques and model comparison.**

# **Confusion matrix**

 In the field of machine learning and specifically the matter of applied math classification, a confusion matrix, conjointly called miscalculation matrix, may be a specific table layout that permits mental image of performance associatealgorithmicprogram, usually a supervised learninone(in unsupervised learning it's typically known as an identical matrix). Every row of the matrix represents the instances {in during in a associate exceedingly in a very} expected category whereas every column represents the instances in an actual category (or vice versa). The name stems from the actual fact that it makes it simple to check if the system is confusing 2categories (i.e. ordinarily mislabeling one as another).It is a special quite contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in each dimensions (each combination of dimension and sophistication may be a variable within the contingency table).

 Where,

**Accuracy**- It is the proportion of total number of prediction that were correct.

**Precision**- It is positive cases that were correctly predicted.

**Negative Predictive Value**-It is the proportion of the negative cases that were correctly identified.

**Sensitivity**-It is the proportion of actual positive case that were correctly identified.

**Specificity**- It is the proportion of actual negative cases that were correctly identified.

For example confusion matrix of Naive Bayes Theorem in **Figure 3** and **Figure 4.**



**Figure 2- Confusion Matrix Diagram**

**Figure 3- Confusion matrix for Naive Bayes Theorem**



**Figure 4- Metrics of Naïve Bayes theorem**

 **Figure 5-Metrics of Logistic Regression**

Like the Confusion Matrix of Naive Bayes and Logistic regression we also find of Gradient Boosting and Deep Neural network for comparison.

# **Cumulative gain and lift charts.**

 Lift is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model. Cumulative gains and lift charts are visual aids for measuring model performance .Both charts consist of a lift curve and a baseline .The greater the area between the lift curve and the baseline, the better the model.

# **Kolmogorov-Sminrnov chart**

K-S or Kolmogorov-Smirnov chart measures performance of classification models. More accurately, K-S is a measure of the degree of separation between the positive and negative distributions. The K-S is 100 if the scores partition the population into two separate groups in which one group contains all the positives and the other all the negatives. On the other hand, if the model cannot differentiate between positives and negatives, then it is as if the model selects cases randomly from the population. The K-S would be 0. In most classification models the K-S will fall between 0 and 100, and that the higher the value the better the model is at separating the positive from negative cases.

# **Receiving operating characteristic chart**

 The ROC chart is similar to the gain or lift charts in that they provide a means of comparison between classification models. The ROC chart shows false positive rate (1-specificity) on X-axis, the probability of target=1 when its true value is 0, against true positive rate (sensitivity) on Y-axis, the probability of target=1 when its true value is 1. Ideally, the curve will climb quickly toward the top-left meaning the model correctly predicted the cases. The diagonal red line is for a random model





**Figure 6-Cumulative Gain chart of gradient boosting Figure 7-lift chart of logistic regression**

 **Figure 8-K-S chart of Gradient boosting**

**Figure 9-ROC chart of Naive Bayes Theorem**

 For every Machine Learning Algorithm mentioned in this paper the above mentioned evaluation methods and charts are used for better understanding and which is better algorithm for using.

# **V. Web deployment.**

A RESTful API is an application program interface (API) that uses HTTP re- quests to GET, PUT, POST and DELETE data. Flask is a micro web framework written in Python. It is classified as a micro frame- work because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. It is relatively easy to set up a website on Flask using Jinja2 templating. Heroku is a cloud platform as a service (PaaS) supporting several programming languages. We create the simple flask app. The flask app consists of 2 main components: the python app (app.py) and the HTML templates. While we can return HTML code from the python file itself, it would be cumbersome to code entire HTML as a string in the python file. This will help us to use the prediction model for new data too.

# **Conclusion.**

 In this paper, we have successfully used different Machine Learning Algorithm, for bank loan default prediction. The task was to predict whether the loan applicant will be a defaulter or will be able to pay the loan.it was implemented in python programming language. While carrying out this process we have found that Gradient Boosting is more accurate followed by Deep Learning for the accuracy. We are also able to know pros and cons of the algorithms. With the help of python flask we can make it available for prediction of new data. This paper provides an effective basis for loan credit approval in order to identify risky customers from a large number of loan applicants using predictive modeling.

# **References**

1. Home Credit Data sources : h[ttps://www.kaggle.com/c/home-credit-default-](http://www.kaggle.com/c/home-credit-default-risk)risk.
2. Home Credit Default Risk by Jay Borkar at Kaggle Competition.
3. Sudhamathy G and Jothi Venkateswaran ” Analytics Using R for Predicting Credit Defaulters”,IEEE international conference on advances in computer applications (ICACA), 978-1-5090-3770-4, 2016.
4. M. Sudhakar, and C.V.K. Reddy, “Two Step Credit Risk Assessment Model For Retail Bank Loan Applications Using Decision Tree Data Mining Technique”, International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), vol. 5, no.3, pp. 705-718, 2016.
5. Dileep B. Desai, Dr. R.V.Kulkarni “A Review: Application of Data Mining Tools in CRM for Selected Banks”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (2) , 2013, 199 – 201
6. Consumer Credit Risk models via Machine Learning Algorithms by AE Khandani, AJ Kim, AW Lo, https://dspace.mit.edu/openaccess-disseminate/1721.1/66301
7. ManjeetKumar, Vishesh Goel, Tarun Jain, Sahil Singhal, DR. Lalit Mohan Goel. (2018). Neural Network Approach To Loan Default Prediction, International Research Journal of Engineering and Technology (IRJET) , p-ISSN: 2395-0072
8. Li Ying. (2018). Research on bank credit defaultprediction based on data mining algorithm. The International Journal of Social Science and Humanities Invention 5(06): 4820-4820, ISSN: 2349-2031.
9. Steenackers, A., & Goovaerts, M. J. (1989). A credit scoring model for personal loans. Insurance: Mathematics and Economics, 8(1), 31 -34. Ali Bagherpour. (2006). Predicting Mortgage Loan Default with Machine Learning Methods.
10. Predicting Bank Loan Default with Extreme Gradient Boosting Rising Odegua Department of Computer Science Ambrose Alli University Ekpoma, Edo state, Nigeria.
11. Machine learning application in loan default prediction by Abhishek Tiwari, Tata Consultancy Services. Novateur Publications JournalNX- A Multidisciplinary Peer Reviewed Journal ISSN No: 2581 - 4230 Volume 4, Issue 5, May -2018