IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Customer Loan Approval using Semi-Supervised Learning

[1]Mohd Abdul Ahad Usama, [2]Mohd Younus, [3]Shaik Sultan

[1]Student, [2]Student, [3]Student

[1]Computer Science and Engineering,

[1]ISL Engineering College, Hyderabad, India

***Abstract:*** Machine learning is playing an important role in our developing world. Many applications are manages and controlled by machine learning. By the utilization of historical data there are possibilities to soothsay the future. Even-though there are presently a number of researchers working yet, the improving the performances and the exactness of the algorithms. Here, performance analysis is the key depends upon the algorithm selected for the data to focus on the work, where it can provide precision in predicting loan approval status.

***Index Terms*** **- Machine Learning, Algorithms, Logistic Regression, Random.**

## I. INTRODUCTION

Machine Learning provides with the efficient and accurate useage of the data. Once in a while in the wake of review of the information.[5] In Machine Learning some tasks are assigned to a computer and it is said that the machine provides the improve in machine's measurable performance gains more and more experience by executing these tasks which it has learnt from its experience. [6] With the plenitude of data set accessible, the interest for machine learning is in ascend. Numerous businesses from drug to military apply machine deciphering how to extricate applicable data. The present inundating worldview for ML is to run a ML calculation on an offered dataset to engender a model. The model is then connected, all things considered, and the assignments are executed. And thus as a result it becomes valid for supervised as well as unsupervised learning.

## II. MACHINE LEARNING

### 2.1 SUPERVISED LEARNING

[3]Here, as the name suggests that this is a supervised method, where teaching and training performed machine. The data-set set contains observations along with features and the target value to train the machine . This is called learning from models that is gaining. There must be a enough data for the machine to train generally composed as (xi,ti), where xi are the sources(features), the objectives are ti, injuctively authorized or ranging from 1 to some limit N.

### 2.1.1REGRESSION

[1]Regression is a probable way to deal with demonstration of the cognation between a scalar reaction (or dependent variable) and at least one illustrative factors (or independent). It can be utilized to model the linear as well as non-linear regression based on the target values which may be either continuous or discrete.

### 2.1.2CLASSIFICATION

It is utilized to previse clear-cut class-names and classifies information/data dependent on preparing of the set and class marks and it may be very well be utilized for authoritatively mandating recently accessible data. [8] Grouping system is a perceived strategy in incipient circumstances for more than once settling on such culls in incipient circumstances. And here, in the event where we expect that issue is a worry in the development of a methodology that will be related to a proceeding with succession of cases in which each incipient case must be assigned to one of an abundance of pre-characterized classes predicated on the optically canvassed highlights of the data. Engendering the arranging system from a plethora of information for which the correct classes are referred to progress of time is denominated as example Pattern apperception and supervised learning. For instance, an arrangement is incorporate fundamentally, relegating people to credit status predicated on their details and other individual data, or to predict the illness of patient based on the symptoms. The most important thing is that the decisions are to be taken depending on the classification of the data after performing the Supervised Learning.

### 2.2UNSUPERVISED LEARNING

[1]Unsupervised learning is marginally more harder on the grounds that here the model needs to be deciphered on the procedure of performance designated assignments without revealing to it how to perform. As a result, this ends up genuinely hard to state what the aim of this learning is. Unsupervised learning can be dealt in by two methods.

[5] The principal approach is to show the operator not by giving rapid orders or direct orders, but rather by giving a type of reward framework to demonstrate the achievement. In decision making problems this kind of learning is very well utilized since in decision making problems the aim is to settle to a choice and classify the problem.[6] Unsupervised learning works to find the natural partitions patterns.

And second methodology is kenned as clustering, where there is no reward framework, it discovers likenesses in the training data. There is conventional suspicion that the cluster will coordinate sensibly well with an instinctive order.

### 2.3REINFORCEMENT LEARNING(RL)

[8]This method lies in the middle of Supervised learning and Unsupervised learning. It needs to investigate and experiment with various potential outcomes until the solution found is right. RL some point called learning with a commentator as a result of this screen scores the appropriate response, however does not recommended upgrades.[10] The data to be worked upon is in two parts classified as one part labelled and another part as un-labelled data. Here the machine trains it-self using and the labelled data and implements the knowledge on the unlabelled data. The target values here are the labels.

Here, the algorithms or the agents are focused to maximize the outcome as reward and minimize the penalty. Unlike the supervised learning where the machine is trained based on the answer but here in reinforcement learning there is no answer instead it learns from its own experiences.

## III. ALGORITHMS

As there is a lot of data out in the world and that too in different formats or may be differently classified based on the features. So there comes an issue since one algorithm might not be capable of providing the outcome, so selecting of an algorithm is the important part in machine learning since there are a lot of different kinds of data. [12]For an instance there are two kinds of data classified as categorical and numerical so the algorithms suitable for particular type of data might be the worst case for the other kind of data. Another example is continuous and discontinuous data. So the selection of an algorithm mainly depends on the features and their values, so as to provide with best algorithm. Some of the simple and most commonly used and the algorithms tested in our work are explained in brief.

### 3.1NAIVE BAYES

As being one of the simple and short algorithms its performance is quick. With the help from Bayes Hypothesis it can foresee the unknown data-set belonging to the particular class. So this strategy is dependent on the Bayes Theorem completely. [2]This methodology is based on conditional probability.

The most advantageous property of this algorithm is that it works well with small data sets as when compared to other algorithms. It works with continuous data discrete data can handle binary and multi-class data as well.

### 3.2LOGISTIC REGRESSION

In logistic Regression the result or the target values encompasses categorized data. Since the target values are classified, this algorithm predicts the class(target value) for an unknown data-point. The target values might have number of classes. [1]Logistic regression is utilized when the reaction variable is clear cut in nature. There ought to be no high relationships among the indicators.

### 3.3LINEAR REGRESSION

Linear Regression is mostly utilized to model the continuous variables that is when the target values are continuous, unlike in Logistic regression where the target values are classified into different classes, here the values are continuous. For observation the other features are plotted against the target value on a graph so if there is a linear increase or decrease in the graph indicates that this algorithm was good for the specific data-set. The features to be plotted against the target value are such that the target value being the independent variable lies on the x-axis, where-as the other features lie on the y-axis indicating the dependent variables.

### 3.4 DECISION TREE

[7]Decision tree is the most amazing and prevalent instrument for arrangement and expectation. The decision tree starts with a single node(leaf) and this leaf is assigned a label over the vote of all the other labels. The Decision tree looks like a flowchart in a tree format. And once the node has a child node which is a decision and each branch, points a result and the each leaf which are terminals holds the result to the root node and are labelled with a class-name. Each level of nodes chosen and construction by calculating correlation between the class and individual features. The error in the data calculated by entropy of the class and information gain of the feature.

### 3.5RANDOM FOREST

[4]As the name of this algorithm name suggests that it is a forest, since a forest has huge number of trees, it's the same case here random forest consists of many decision trees. This is an ensemble algorithm since it encompasses of many decision trees. If tree becomes a target class all the other classes of all different features are directly or in-directly child to the tree with the target class. And the key here in the creation of the random forest is that all the child trees must be strongly associated to each other. [11] The most fascinating thing about the random forest is the habits in which that it makes arbitrariness from a standard dataset. The prediction in Random Forest is done by the result of voting of the trees in the random forest and the majority of the votes decide the class for the data-point in the data-set.

## IV.RELATED WORK

The work we have done is on Machine Learning domain which is building a model to predict the customer's loan request getting approved based on the data provided to the machine to train itself and then implementing the knowledge on other observations to classify the customer whether his/her loan request gets approved or not.

Since we have done some research on the algorithms, just briefed above. Using those algorithms we made models to justify the best algorithm providing us more accuracy and less error rate.

**4.1Data-Set**

The data-set we used for our study on was obtained from a public club repository loan data. The retrieved data consists of the customers with and without defaulted loans. The data-set obtained was a large containing a huge number of observations/records and we retrieved only three-thousand observations from the data set to make our study simple. The data-set has independent features and one dependent feature. The independent features which were not much in contribution on the dependent feature were dropped and to make the study a bit more simple only important were selected. The dependent feature is 'Loan_Status'. This dependent feature has two different classes namely paid and not-paid. So here we have the target value in classes so it comes under categorized data.

**4.2Building The Model**

The model to be built requires the algorithms which are good for categorized data as we have to predict the class for a customer from the three classes. We used random forest as a classifier for our study and compared it with logistic regression since both work well for categorized data.

The model was built using python language and the respective libraries were used. The data-set before classification was to be cleaned. So we checked for null values and dropped some records where we were not able to fill in the values with proper mean/averages method. The data set is checked for duplicate values and hence dropping is done. After cleaning the data-set we made categories based on the features consisting more outliers as in case for income the values has which is varied for customer to customer and the categories were made for the customers based on the income as when the income is greater than 1,00,000 is named as 'high', income of 30,000 to 1,00,000 is named as 'medium' and income below than 30,000 belongs to 'low' class. After the classification based on income, the values are transformed into numerical as we are going to use logistic regression which takes numerical data to train. So the features with non-numerical categories are transformed to numerical but they remain as categories only.

**4.2.1Steps used are presented below:**

Step-1: The data-set is cleaned, finding the missing values, filling the missing values with mean values or standard deviation methods and dropping some records.
Step-2: After cleaning the data duplicates are checked and removed for data redundancy.
Step-3: The data transformation is done based on income as well as on non-numerical to numerical
1: if income>=100000 grouped in 'high'
2: if income<100000 and >=30000 grouped in 'medium'
3: and lastly if income <30000 grouped in 'low'
Step-4: The data set is split into train and test data.
Step-5: Initialization for the algorithms (random forest and logistic regression)
Step-6: Training is done after initialization.
Step-7: Now the test data-set is tested, this model is created and can be deployed directly by users.

**4.2.2Data Cleaning:**

The data might be dirty, in the sense that some of the record's values might not be noisy missing. So the process for cleaning is called data-cleaning. This step of data pre-processing increases the data quality.
The missing values can be processed by:
1. Ignoring the tuple.

2. Using some global constant to fill on the value.

3. Using mean or median to fill the value.

4. Filling in most probable value (find using the decision tree).

**4.2.3 Data Reduction:**

1. Dimensionality reduction is the process of reducing the number of attributes under consideration.

2. Numerosity reduction is the process where original data values are replaced by other smaller forms of data to make the processing easy.

**4.2.4 Outliers:**

Some data which do not comply with other data-points in the data set are the called outliers. Many times the outliers decrease the accuracy. So to make it efficient the data feature is classified into class to properly categorize the data.

**4.3 Performance, using the Random Forest Classifier**

| S. No. | Percentage Split | Accuracy | Correlation | Error Rate |
|---|---|---|---|---|
| 1. | 80%,20% | 82.16 | 0.97 | 0.256 |
| 2. | 70%,30% | 81.35 | 0.95 | 0.276 |
| 3. | 60%,40% | 76.23 | 0.89 | 0.270 |

**4.4Performance, using the Logistic Regression**

| S. No. | Percentage Split | Accuracy | Correlation | Error Rate |
|---|---|---|---|---|
| 1. | 80%,20% | 82.13 | 0.99 | 0.23 |
| 2. | 70%,30% | 80.89 | 0.92 | 0.28 |
| 3. | 60%,40% | 70.63 | 0.86 | 0.298 |

## V CONCLUSION

The study has produced the best known and presented in brief. The references taken were at the time of the study period. Our work was to deliver the knowledge gained in simple and best way possible. We tried our best to make this study and our work on building the model as simple as possible as Data-Science being one of the top most domains. The references taken from books and published papers made our study more easy and helped us to make it more simple.

## REFERENCES

[1] R. KarthiBan, "A review on Machine Learning classification technique for Bank Loan Approval", published IEE, 2019
[2] Susmita Ray,"A quick review of ML algorithms", published IEEE 2019
[3] https://www.geeksforgeeks.org/supervised-unsupervised-learning/
[4] Shai Shalev-Schwartz "Understanding Machine Learning : From theory to Algorithms," published 2014 by Cambridge University Press.
[5] Stephen Marsland, "Machine Learning – An Algorithmic Perspective", Second Edition, Chapman and Hall/CRC Machine Learning and Pattern Recognition Series, 2014
[6] Ethem Alpaydin, "Introduction to Machine Learning 3e (Adaptive Computation and Machine Learning Series)", Third Edition, MIT Press, 2014
[7] Tom M Mitchell, "Machine Learning", First Edition, McGraw Hill Education, 2013.
[8] Peter Flach, "Machine Learning the Art Science of algorithms that makes sense data", First Edition, Cambridge university Press. 2012
[9] Leo Breiman, Jerome H. Friedman, Richard A.Olshen and Charles J. Stone, "Classification and Regression Trees," Wadsworth Publishing ,1983
[10] Xiaojin Zhu and Andrew B. Goldberg, "Introduction to Semi-Supervised Learning," 2009
[11] S.B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", Informatica 31 (2007) 249-268
[12] Nils J. Nilsson, "Introduction to Machine Learning",Robotics Laboratory,Stanford University 1998.