ISSN: 0976-3031

International Journal of Recent Scientific Research

Impact factor: 5.114

A MACHINE LEARNING APPROACH TO PROFILING

Rynah Rodrigues

Volume: 6 Issue: 10

THE PUBLICATION OF
INTERNATIONAL JOURNAL OF RECENT SCIENTIFIC RESEARCH
http://www.recentscientific.com
E-mail: recentscientific@gmail.com

Available Online at http://www.recentscientific.com

ISSN: 0976-3031

International Journal of Recent Scientific Research
Vol. 6, Issue, 10, pp. 6826-6829, October, 2015

International Journal of Recent Scientific Research

RESEARCH ARTICLE

A MACHINE LEARNING APPROACH TO PROFILING

Rynah Rodrigues

Student, Bachelor of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Mumbai, India

ARTICLE INFO

Article History:

Received 05thJuly, 2015
Received in revised form 08thAugust, 2015
Accepted 10th September, 2015
Published online 28st October, 2015

Key words:

banks, loans, machine learning, profiling

ABSTRACT

Human beings are quite complex. Every individual is different and various set of parameters need to be evaluated according to the background of the person in order to prepare a profile of any particular individual. However, some patterns emerge and equipped with machine learning some predictions can be construed. This paper presents profiling of candidates applying for Home Loans using various machine learning algorithms and techniques. This will enable bank officials to better understand their customers, thereby reducing the ratio of loan defaults. The myriad of data present with the bank which relate to past records of home loan candidates have been used by the machine learning algorithm in order to learn and produce a profiling output. Candidates are then placed into categories and a further investigation can be carried out depending on which category a candidate is allocated to.

INTRODUCTION

Banks all over the world offer loans to their customers, which benefit the banks through interest rates. However, the issue of customers defaulting the loan is quite common. Many banks face major problems due to these loan defaults. It thus becomes highly imperative that every applicant is screened thoroughly before the loan has been issued. There are many factors which a loan officer will need to take into consideration when we consider home loans. Every applicant for a home loan is different. They all come from varied backgrounds and every case as a whole is unique.

However, a loan officer with experience gets better and better at the screening process using past experiences, relating some factors from past cases to make present decisions. This same concept can be applied in making a computer improve itself with experience. This is basically the idea behind machine learning. The following can be achieved by using machine learning as opposed to using a human to carry out the same function:

- Eliminates prejudices ensuring that the results are unbiased
- Larger amount of data can be processed.
- Results can be produced at a faster rate.
- More accurate results are obtained.

Banks contain enormous data and past records related to the domain of home loans. This large amount of information could be studied in order to improve the analysis of the loan application process. Loan officers would find it difficult to retain and work effectively analyzing all the customers' data. This is where Machine Learning could aid in developing a computer based system to carry out this similar process and update itself regularly with changes in data. However, the methodology used is limited to the fact that the results obtained from this method is only a predictive result and not a guaranteed one.

Machine Learning

Machine learning is a type of artificial intelligence that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data. In machine learning the data is processed to extract patterns and modify the program according to these patterns. [1]

Machine Learning algorithms can broadly be classified into two categories namely, Supervised Learning and Unsupervised Learning.

Corresponding author:* **Rynah Rodrigues
Student, Bachelor of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Mumbai, India

Supervised Learning

In this type of learning, the data set is provided in such a way that it gives the correct answer. For every example, the data set provides the right answer and the task of the algorithm is just to produce more right answers given a definite value for an attribute. The input to the algorithm is a training set from which the algorithm attempts to learn. The output can be of two types which are Regression or Classification. Regression produces continuous valued output i.e there can be several different values within the range of the training set. Classification on the other hand categorizes the output into a discrete set of values i.e the output will fall in any one category defined by the problem at hand. [3]

Unsupervised Learning

Unlike supervised learning where we are explicitly told what the correct answer is for an example i.e each output falls under some kind of label, in unsupervised learning, the output data does not fall under any label. Given a data set we are asked to find some structure in the data. The unsupervised learning algorithm breaks the data into clusters. The algorithm places the output in one of the clusters. [3]

METHODOLOGY

While developing a machine learning model, it is more effective when several different machine learning algorithms have been considered. The idea is very simple - if you gather predictions from dozens of Machine Learning models then the average score will be better than the best single prediction [2]. An alternative to this could be to use fewer or a single Machine Learning algorithm but instead of combining all the features in a single model, a subset of the features could be processed separately and the predictive results obtained from the various subsets. Then all the features are combined together if requiredin order to get a complete overview of the problem. The benefits of designing the system in this manner especially when considering a home loan application problem is that when smaller groups of features are processed, it gives a chance to track where the applicant fails in their eligibility. This will enable an applicant to make an attempt to improve upon a particular attribute and also to mitigate the risk factor. Placing all the features into a single model would just produce the final result not providing a detailed result as to what the applicant lacks in order to satisfy the eligibility criteria.

According to the problem definition in this case, a combination of various supervised learning algorithms is best suited to produce effective and best predictive results. As banks possess a lot of data concerning their past applicants for home loans, this abundant data could be used as the training data so that the algorithm can learn from it.

Data Categorization

The first step while developing the system is categorization of the data. This is done by getting an idea of the key aspects of the problem and then segregating the data accordingly. In the case of home loans, the following higher level of classification shown in Fig 1. best suits the purpose. There are many features that would need to be considered, Fig 1. includes only the key aspects [4].

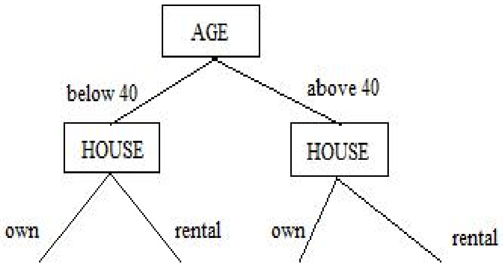

Fig 1 Data Categorization

Features such as default ratio (number of loan repaid/number of loans taken), number of dependents of the applicant, collateral, employment etc are additional features that make up the categorization. Categorization makes it possible to obtain a detailed analysis by training the data for smaller aspects of the problem. Where to put an end to the classification is subjective and will vary depending on the intended purpose or goal.

Combining Features

Once the data has been categorized, the next step is to determine what information needs to be extracted from this data. As discussed early, two or more features could be combined at a time to obtain preliminary profiles of a home loan applicant. Then all the features could be combined to obtain an overall profile of the applicant. This process enables an applicant to track what they lack and in which areas they need to improve.

Extraction of information from home loan data begins with an understanding of what we want to know. For instance an applicant might want to know based on his monthly installment of the home loan and 60 percent of his take-home salary if he needs a collateral or not.

Another could be based on the applicant’s loan amount, how much can his default ratio be. If the applicant has more than the predicted value then he might not be eligible for a loan. Similarly, many aspects of the loan applicant can be processed in order to obtain a detailed analysis. The final step is to combine all the features in order to obtain a model which provides a predictive result as whether an applicant is a high risk or a low risk applicant.

RESULTS AND DISCUSSION

Using Matlab the training data is fed to the algorithm and predictive results are obtained. Based on the selection of the features and the manner in which the data presents itself a Machine Learning algorithm is chosen that best delivers the results.

Result of a Combination of features

Considering the example discussed earlier of combining the features of applicant’s loan amount and default ratio the following graph was obtained when the data was fed to the algorithm. Fig 2. presents a scatter plot of the past success records of applicant above 40, living in their own houses being plotted and a line which bests fits the plotted data has been obtained by a linear regression algorithm [6].

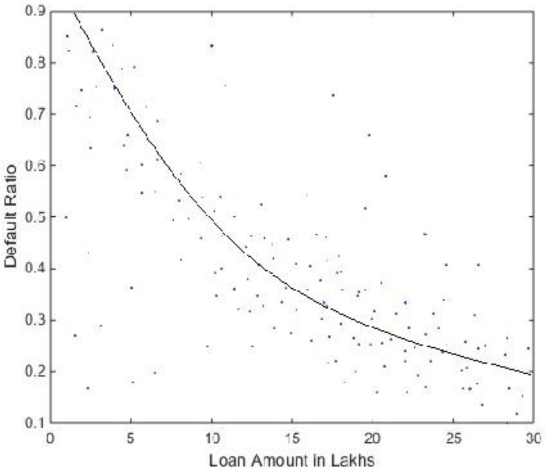

Fig 2 Graph to determine default ratio eligibility

Using the graph in Fig 2. the following test cases were processed for applicants in order to determine if their default ratio satisfied the eligibility criteria.

If the applicant's default ratio is more than the predicted value then the applicant has failed the default ratio eligibility aspect. In Table 1. if the applicant's status is 1 then the applicant has passed this aspect of the eligibility else if 0 then the applicant has failed.

Table 1 Test Cases

Loan Amount	Applicant's Default Ratio	Predicted Default Ratio	Applicant's Status
23 lakhs	0.68	0.32	0
9 lakhs	0.52	0.55	1
25 lakhs	0.10	0.25	1
30 lakhs	0.34	0.20	0
5 lakhs	0.49	0.72	1

From the table above an applicant will be able to obtain a detailed analysis of the home loan application process. In a similar way other machine learning algorithms and models can be used based on the requirements and the way in which the data is present.

Combining the preliminary predictive results

Based on the results obtained when a subset of the features were combined for analysis, an overall probability can be obtained about the applicant's eligibility for a home loan. A classification model can be built for the same using a logistic regression algorithm.

For the purpose of visualization on a 2 dimensional plane and to acquire a more profound understanding of how a classification model using logistic regression would look, Fig 3. presents two features from the problem set and uses it to classify applicants into two categories namely accepted and rejected.

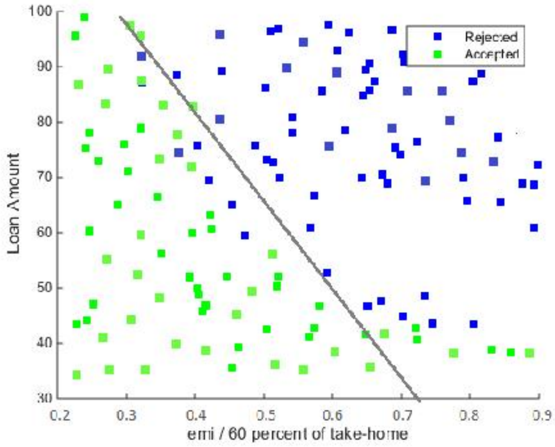

Fig 3 Classification Model

The past records of successful and defaulted applicants for the home loans is used as a training set for logistic regression. For every training example an applicant's loan amount and the ratio of his estimated monthly installment and 60 percent of his take-home salary is present along with the status which indicates if the applicant had successfully paid the loan or defaulted on it. The line which passes through the data set is known as the decision boundary. This separates the applicants who paid the loans from the ones who defaulted on the loans. Using this we can predict in which category an applicant who is currently applying for a home loan belongs.[5] and [6]

CONCLUSION

The major contribution of this research is to profile applicants using machine learning algorithm. It not only helps predict the profile of an applicant but also provides a detailed analysis so that the entire process is transparent providing applicants with an idea of how they can improve if required.

Profiling of the applicants for home loans will enable the banking industry to improve on their analysis of the process thereby mitigating the risk associated with home loans. It also helps in speeding up the process. Experience is the best teacher and building upon this the use of Machine Learning to this process is quite effective. The loads of past banking records used to train the machine will help reduce errors. As the machine is unbiased and without prejudice the results obtained are guaranteed to be fair as opposed to some of the decisions made by humans in this field. This process could also help train novices in the field of banking and could help them with the experience they require in this domain.

References

1. whatis.techtarget.com/definition/machine-learning
2. datalab.lu, Applying machine learning to peer-to-peer lending.
3. www.aihorizon.com/essays/generalai/supervised_unsupervised_machine_learning.htm

4. Tetsuo Tami and Masayuki Fujita, Development of an Expert System for Credit Card Application Assessment. Mitsubishi Research Institute, Inc.
5. Taha Zaghdoudi, Bank Failure Prediction with Logistic Regression. International Journal of Economics and Financial Issues.
6. Maria Aparecida Gouvêa and Eric Bacconi Gonçalves, Credit Risk Analysis Applying Logistic Regression, Neural Networks and Genetic Algorithms Models. POMS 18th Annual Conference.

****

How to cite this article:

Rynah Rodrigues *et al*.2015, A Machine Learning Approach To Profiling. *Int J Recent Sci Res* Vol. 6, Issue, 10, pp. 6826-6829.

International Journal of Recent Scientific Research

ISSN 0976-3031

9 770976 303009