Loan Sanctioning Prediction System

Aditi Kacheria, Nidhi Shivakumar, Shreya Sawkar, Archana Gupta

Abstract*— People operating in banks face lots of issues which involve approval of a loan. In the 21st century, people often rely on technology to tackle such issues. This paper proposes a loan sanctioning system which determines whether or not a loan should be given to a person, based on certain attributes. In spite of banks following stringent rules and regulations and conducting meticulous background checks while sanctioning a loan and keeping in mind the probability of the person's ability to return the loan, often such situations are faced where in, the person is unable to repay the loan that has been given to him. In this paper, the system that we propose for the bankers will help them predict the credible customers who have applied for loan, thereby improving the chances of their loans being repaid in time. This classification is done using Naïve Bayesian algorithm. In order to improve the classification accuracy, the quality of the data is improved before classifying it by using K-NN and Binning algorithms. This system uses these algorithms in order to yield a better efficiency so as to reduce the possibility of such a problem. The proposed system additionally facilitates self-confirmation regarding the same for the commoner.

***Index Terms*— Binning, Data mining, K-NN, Naïve Bayesian.**

I. INTRODUCTION

Data mining is actually the invention of valuable information and patterns from large chunks of accessible knowledge [7]. It is widely used by companies in order to extract useful information from large amounts of data. Conversion of raw data into useful information thereby analysing the data for relationships that have not previously been discovered is done using data mining [9]. It acts as a powerful tool for the implementation of artificial intelligence which is an emerging trend in various sectors. Different areas where it is used are- Medical care, Finance, Economics, Telecommunication, Sales, Marketing, Recommendation Analysis and Fraud Detection [7]. Using certain techniques in data mining, specific results from a large data set can be found very quickly and easily. Due to high competition within the business field, it's essential to think about the client relationship management of the enterprise. Here we analyse the large volume of client knowledge and classify them based on the client behaviours and prediction. Client relationship management is especially employed in sales prediction and banking areas [8]. We propose to build a system in order to predict whether or not a loan can be sanctioned to a particular person by making use of data mining strategies. Our system focuses on being self-used by a commoner, thus saving the tedious procedure of visiting banks in order to find out whether or not a loan will be sanctioned to him. Naïve Bayes is one of the successful data mining techniques used, based on which we will be building our system.

Revised Version Manuscript Received on August 06, 2016.

Aditi Kacheria, Student, Department of Computer Engineering, K. J. Somaiya College of Engineering, Mumbai, India.

Nidhi Shivakumar, Student, Department of Computer Engineering, K. J. Somaiya College of Engineering, Mumbai, India.

Shreya Sawkar, Student, Department of Computer Engineering, K. J. Somaiya College of Engineering, Mumbai, India.

Prof. Archana Gupta, Assistant Professor, Department of Computer Engineering, K. J. Somaiya College of Engineering, Mumbai, India.

II. OVERVIEW OF THE PROPOSED MODEL

Attributes:

Our loan sanctioning process predicts whether the loan amount requested by the customer will be granted to him or not.

To arrive at the conclusion we use the following attributes [11]:

i. Age
ii. Profession
iii. Total Income
iv. Existing Loan
v. Loan Tenure
vi. Loan Amount
vii. Loan Approved

Taking into account the values of these attributes, the system uses the Naive Bayes classifier to classify the given case into one of the pre-defined classes. Based on this classification we give the result as Yes (suggesting that loan will be sanctioned) or No (loan will not be sanctioned).

Components of the System:

A. Pre-Processing
B. Classification
C. Database Updation

A. Pre-Processing:

i. Data Completion: In the real-world database, there are often missing values in the dataset due to improper data entry or other data entry problems that can lead to issues in data analysis, thereby affecting the accuracy of the system [1]. In order to overcoming this issue of missing values, K-NN algorithm is used.

ii. Data Refining: Data sets often have outliers that are nothing but noisy records present in the data [7], [11]. These outliers need to be removed in order to further increase the accuracy of the system, which is done using the Binning algorithm.

B. Classification:

The final result to be obtained is whether or not the loan will be sanctioned to the user. The data is bifurcated into two pre-defined classes i.e. Yes or No, depending on whether or not the loan is sanctioned, respectively. The Naive Bayesian classification technique is used to give a result based on the data, to the user as Yes (the loan will be sanctioned) or No (the loan will not be sanctioned).The

authors of the accepted manuscripts will be given a copyright form and the form should accompany your final submission.

C. Database Updation

We add this newly found data i.e. the data provided by the user, to our existing data set so that the database is more accurate and up to date.

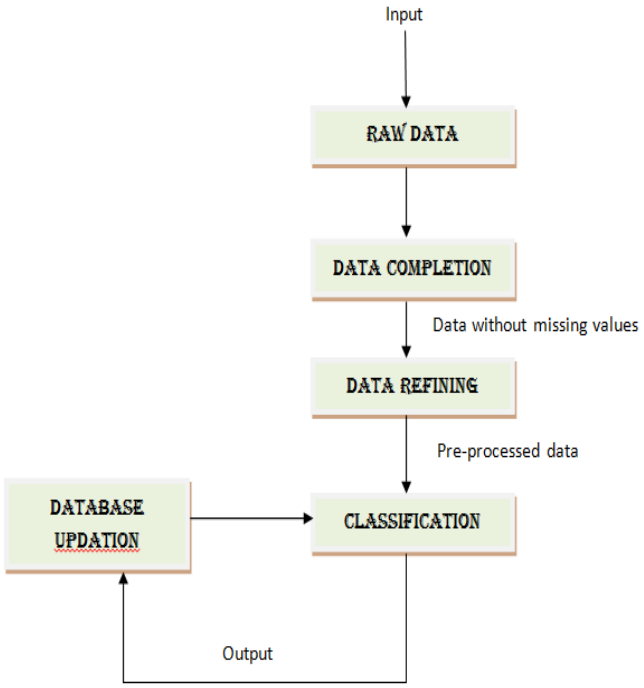

Fig. 1: Flowchart of the Proposed Loan Sanctioning Prediction System

III. PRE-PROCESSING METHODS

A. Data Completion:

Data entry problems can lead to the presence of certain data with missing values that can cause complications in the data analysis [7]. To overcome the problem of missing values, the system uses the K-NN algorithm. The algorithm uses the Euclidean formula to calculate the distances to all neighbours.

After sorting these distances, it determines the k-nearest neighbours. The value of this k is dependent on the application [7]. The next step of this algorithm is where the chosen neighbours vote for the value that should be filled in the missing space. After determining the value with the maximum votes, the system fills the missing space. The accuracy of this algorithm is good but the speed decreases with the increase in the size of the data set. Since the process of loan sanctioning does not demand immediate response, the speed of this algorithm is sufficient. Hence the K-NN algorithm is suitable for the proposed system.

B. Data Refining:

The data set may contain exceptions, inconsistent data or outliers. To remove this and make the values of the data set consistent, the system uses the Binning algorithm. The output from the K-NN algorithm is given directly to the binning algorithm. The Binning algorithm is a data smoothing technique. It first takes the data set and sorts it. After sorting, the data set is distributed into different bins of fixed sizes. Size of the bins is dependent on the application. The data stored inside the bin is replaced by the mean of each bin. This makes the data smooth and removes the outliers.

IV. CLASSIFICATION TECHNIQUE

A. Naive Bayesian algorithm:

In Artificial Intelligence, there are two classes - Supervised Learning and Unsupervised Learning. The Naïve Bayesian algorithm falls under the category of Supervised Learning [9]. Supervised Learning, also called Machine Learning, is data-driven. These algorithms calculate the probability and make a decision based on the dataset. Probability is a measure of how likely it is that the event will occur [3]. Naive Bayesian classifier depends on Bayes' theorem that works on probabilistic statistical classifier. The Naive Bayesian Classifier technique is most appropriate to use when the data sets are large and inputs are multi-dimensional, since it is robust and fast [3].Although, Naive Bayesian Technique is easy and does not use any complicated iterative parameter estimation, the results are reliable. Naïve Bayes classifier uses the strategy of parallelism [9]. In this loan sanctioning system, the Naive Bayesian technique is used to classify the query into two parts i.e. yes or no.

Yes indicates that the person's loan will be sanctioned whereas, No indicates that the loan will not be sanctioned.
The Bayes theorem is as follows:
The target can be one of discrete values: t1, t2, …, tn.
Let T= {t1, t2, …, tn} and P be the probability.
The set of attributes are {A1, A2,…, An}

$$T = \arg\max_t P(T = t \mid A_1 \ldots A_n)$$

$$= \arg\max_t P(A_1 \ldots A_n \mid T = t) * P(T = t)$$

$$P(A_1 \ldots A_n \mid T) = P(A_1 \ldots A_{n-1} \mid A_n T) P(A_n \mid T) =$$

$$P(A_1 \ldots A_{n-2} \mid A_{n-1} A_n T) P(A_1 \ldots A_{n-1} \mid A_n T) P(A_n \mid T) =$$

$$= \prod_i P(A_i \mid A_{i+1} \ldots A_n T)$$

$$T = \arg\max_t P(T = t) * \prod_{i=1}^{n} P(A_i \mid T = t)$$

$$P(A_i \mid A_{i+1} \ldots A_n T) = P(A_i \mid T)$$

$$P(A_1 \ldots A_n \mid T) = \prod_i P(A_i \mid T)$$

$$T = \arg\max_t P(T = t \mid A_1 \ldots A_n) = \arg\max_t P(A_1 \ldots A_n \mid T = t) * P(T = t)$$

$$T = \arg\max_t P(T = t \mid A_1 \ldots A_n) = \arg\max_t P(A_1 \ldots A_n \mid T = t) * P(T = t)$$

$$T = \arg\max_t P(T = t \mid A_1 \ldots A_n) = \arg\max_t P(A_1 \ldots A_n \mid T = t) * P(T = t)$$

Here, the Target is the final result that is produced i.e. Yes or No. Using this algorithm, the probability is calculated for estimating the final result. The probability for both, loan being sanctioned as well as loan not being sanctioned is calculated based on the values of all the attributes. The higher probability is generated as the result.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

V. EXAMPLE

Table 1: Sample Dataset

Age	Profession	Total Income(in lakhs/year)	Existing loans	Loan Tenure(in years)	Loan Amount(in lakhs)	Loan Approved
34	Government	15	50000	6	9	Yes
22	Private	6	0	4	4	Yes
45	Business	9	60000	5	8	No
66	Government	12	120000	2.5	20	Yes
22	Business	5.5	330000	6	15	No
44	Private	7	54000	3	5	Yes
37	Government	2	45000	4	5	No
56	Private	5	12800	3	6	No
33	Private	6	450000	6	8	Yes
68	Government	4	48000	2	6	Yes
64	Private	3	33000	5	12	No
27	Business	8	65000	8	5	No
55	Business	10	44000	10	12	Yes
45	Private	15	230000	6	20	Yes
38	Private	14	78000	10	15	No
63	Business	17	200000	3	4.5	Yes
62	Private	13	150000	5	3	Yes
47	Private	15	125000	2.5	4	Yes
52	Government	9	215000	5	6	No
38	Private	12	0	4	3.5	Yes

Consider:

Age<= 50: Young
Age> 50: Old
Total Income<= 5: Low
Total Income> 5 and Total Income<= 12: Medium
Total Income> 12: High
Existing Loans<=100000: Low
Existing Loans> 100000: High
Loan Tenure<= 5: Low
Loan Tenure> 5: High
Loan Amount<= 10: Low
Loan Amount> 10: High

Let us consider that a person whose Profession is Private has an income of Rs. 10 lakh/year. The person is aged 37 years and has an existing loan of Rs. 75000. He applies for a loan of Rs. 4 lakh for tenure of 3 years. Will his loan application get approved?

To find out whether or not his loan will get sanctioned, we have to calculate the probability of it getting sanctioned and not getting sanctioned using Naive Bayesian classification algorithm.

Calculations:

Attributes = (Age=Young, Profession=Private, Total Income=Medium, Existing loans=Low, Loan Tenure=Low, Loan Amount=Low)

Target = Loan Approved = [Yes | No]?

P(Attributes, Loan Approved =Yes) = P(Age=Young| Loan Approved =Yes) * P(Profession=Private | Loan Approved =Yes) * P(Total Income=Medium | Loan Approved =Yes) * P(Existing loans=Low | Loan Approved =Yes) * P(Loan Tenure=Low | Loan Approved =Yes) * P(Loan Amount=Low | Loan Approved =Yes) * P(Loan Approved =Yes)

=7/12 * 7/12 * 6/12 * 6/12 * 8/12 * 8/12 * 12/20 = 0.0227

P(Attributes, Loan Approved =No) = P(Age=Young| Loan Approved =No) * P(Profession=Private | Loan Approved =No) * P(Total Income=Medium | Loan Approved =No) * P(Existing loans=Low | Loan Approved =No) * P(Loan Tenure=Low | Loan Approved =No) * P(Loan Amount=Low | Loan Approved =No) * P(Loan Approved =No)

=5/8 * 3/8 * 4/8 * 6/8 * 5/8 * 5/8 * 8/20 = 0.0137

P (Loan Approved = Yes | Attributes) > P(Loan Approved = No| Attributes)

Therefore, the Naive Bayesian classifier predicts Loan Approved = Yes for the given Attributes, which implies that the loan for the person with the above mentioned attributes will be sanctioned.

VI. CONCLUSION AND FUTURE ENHANCEMENTS

There are many instances where the data set is inconsistent due to missing values and anomalies hence it is important to apply pre-processing algorithms. For dealing with missing values we apply the K-NN algorithm. K-NN is a simple algorithm that stores all available data and classifies new data based on a similarity measure (e.g., distance functions). The binning algorithm is used for removal of these anomalies. These algorithms will improve the efficiency and make the data set more consistent. For predicting whether the loan will be approved or not, the Naïve Bayes approach is used. Naive Bayes algorithm is a classification algorithm with the naive assumption of independence between every pair of features. The combination of K-NN, Binning and Naïve Bayes algorithms gives us the prediction.

To improve the accuracy of the system, a hybrid of Naïve Bayes and K-means can be used.

REFERENCES

1. Nirmala Devi M.; Appavu S.; Swathi U.V., "An amalgam KNN to predict diabetes mellitus", Emerging Trends in Computing, Communication and Nanotechnology (ICECCN), 2013 International Conference on, pages 691 – 695, 25-26 March 2013.
2. S.X. Wu, S.F. Liu, M.Q. Li, "The Method of Data Preprocessing in Grey Information Systems", Control Automation, Robotics and Vision, 2006. ICARCV '06. 9th International Conference on, pages 1-4, 5-8 Dec. 2006.
3. Ranganatha S.; Pooja Raj H.R.; Anusha C.;Vinay S.K., "Medical data mining and analysis for heart disease dataset using classification techniques", Research & Technology in the Coming Decades (CRT 2013), National Conference on Challenges in, pages 1 – 5, 27-28 Sept. 2013.
4. Sudhakar, K.; Manimekalai, Dr. M., "Study of Heart Disease Prediction using Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, ISSN: 2277 128X, pages 1157-1160, January 2014.
5. D.Lavanya; Dr.K.Usha Rani. "Performance Evaluation of Decision Tree Classifiers on Medical Datasets"International Journal of Computer Applications (0975 – 8887), Volume 26– No.4,pages 1-4, July 2011.
6. Dr. K. Usha Rani, "Analysis of Heart Diseases Dataset Using Neural Network Approach", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.5, September 2011.
7. Karthika Jayprakash, Nidhi Kargathra, Pranay Jagtap, Suraj Shridhar and Archana Gupta, "Comparison of Classification Techniques for Heart Health Analysis System", International Journal of Computer Sciences and Engineering(IJCSE), Volume-04, Issue-02, E-ISSN: 2347-2693, pages 92-95, Feb. 2016.
8. Ms. Neethu Baby, Mrs. Priyanka L.T., "Customer Classification And Prediction Based On Data Mining Technique", International Journal of Emerging Technology and Advanced Engineering (IJETAE), Volume 2, Issue 12, ISSN 2250-2459, ISO 9001:2008 Certified Journal, pages 314-318, December 2012.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

9. Rucha Shinde, Sandhya Arjun, Priyanka Patil, Prof. Jaishree Waghmare, "An Intelligent Heart Disease Prediction System Using K-Means Clustering and Naïve Bayes Algorithm", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 6 (1), ISSN: 0975-9646, pages 637-639, 2015.
10. L.Pandeeswari, K.Rajeswari, "K-Means Clustering and Naive Bayes Classifier For Categorization of Diabetes Patients", International Journal of Innovative Science, Engineering & Technology (IJISET), Vol. 2 Issue 1, ISSN 2348 – 7968, pages 179-185, January 2015.
11. Sivasree M S, Rekha Sunny T, "Loan Credibility Prediction System Based on Decision Tree Algorithm", International Journal of Engineering Research & Technology (IJERT), Vol. 4 Issue 09, ISSN: 2278-0181, pages 825-830, September-2015.

AUTHORS PROFILE

Aditi Kacheria is a Final Year student pursuing her Bachelor's in Computer Engineering from K.J. Somaiya College of Engineering, Mumbai, India.

Nidhi Shivakumar is a Final Year student pursuing her Bachelor's in Computer Engineering from K.J. Somaiya College of Engineering, Mumbai, India.

Shreya Sawkar is a Final Year student pursuing her Bachelor's in Computer Engineering from K.J. Somaiya College of Engineering, Mumbai, India.

Prof. Archana Gupta is an Assistant Professor in the Department of Computer Engineering at K.J. Somaiya College of Engineering, Mumbai, India.