Fraud Customer Prediction Based on Bank Loan Data Analysis Using Machine Learning

Ankita Tiwari[1], Akansha Garg[2], Nagresh Kumar[3]
[1,2,3]*Department of Computer Science and Engineering, Meerut Institute of Engineering and Technology, Meerut 250005, U.P., India*

***Abstract* - In our country, there has been a huge demand of personal loans arise from the citizens. There are so many people who are applying for the personal loan from banks as per their needs. But for the banks, it is difficult to detect the fraud customers that which customer will pay their loans & which will not be due to the number of bank frauds is increasing day by day. To prevent this situation, we have explained how to create predictive loan models. In steps, we have shown how to process the raw data, select relevant features, performed data analysis & lastly built a model. In this paper, we have built some supervised learning models which are having higher accuracy score and on the basis of requests we easily determine which transactions to authorize. Classification report having higher f-score, precision and recall is considered as the best model among all the models.**

***Index Terms* - Accuracy score, Classification Report, F-score, Precision and Recall.**

1.INTRODUCTION

This paper is about to detect the fraudulency of the customers that may have chances to do the scam with the banks. In this, we used the learning algorithms to predict the behavior of the customer [4]. Due to the progression of technology and worldwide communication, fraud has been increasing drastically. Though the data mining techniques are used the result is not much accurate to detect these frauds. The most important product of the banking is loan. To convince customers to appeal their loans, banks are trying to estimate the effective business plans. There are two ways to control the fraud, one is fraud prevention and another one is fraud detection. The major objective of prevention is to rule out from the fraudulent activity and authorize transactions [3]. To avoid this situation, banks have to find some techniques to predict customers' behaviors. Machine learning algorithms which are widely used by the banking have a pretty good performance regarding this purpose. The best statistical method is regression to solve the issue of loan frauds.

2. RELATED WORK

Many researchers came to distinguish immediately that the people that are capable to remunerate the amount in a fixed tenure by using data mining techniques. In paper [1]," An Exploratory Data Analysis for Loan Prediction Base on Nature of the Clients", the main objective is to classify nature of the bank customer who wants the loan from the bank. The main conclusion of this paper was that the clients apply loan for debt consolidation and loans was preferred by the majority of loan applicants. Dr. G. Sudhamathy, [6] discussed that, nowadays to reduce their capital loss, risks are arising at a very rapid rate. It is very difficult for the bank to identify which one is defaulter or not because the related data of customers are present in huge amount. Data Mining is a favourable area of data analysis which is useful to extract knowledge from complex data sets. This paper [2] is about the bank loan analysis by Big Data Approach using Hadoop. In this paper, the main objective is to analyze the loan performance and credit risks of the ‚Lending Club' company. This paper used Hadoop approach and for applying this methodology Cloudera software.

3. PROPOSED WORK PLAN

In this paper, we have compared the accuracy of different algorithms and developed a system which can perform early prediction of customers' behaviors with higher accuracy and f-score as well. The main motive of our paper is to minimize the false-positive parameter i.e., Type-1 error. And depending on that accuracy, f-score and false-positive parameters choosing the algorithm which performs best for predicting whether the customer has able to pay the loan or not. By this approach, banks can detect the default behaviors at the earlier stage and control the consistent actions to reduce the possible loss.

3.1 MODEL DIAGRAM

In diagram given below, a systematic procedure is outlined that shows the flow of the research conducted in building the model.

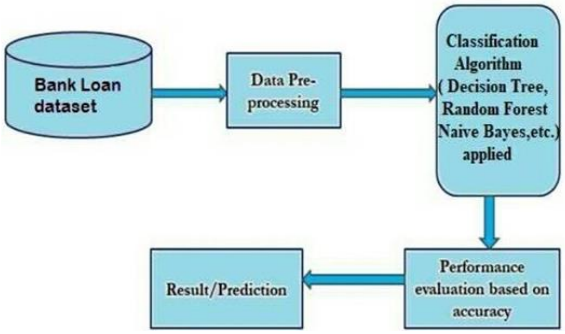

Bank Loan dataset
Data Pre-processing
Classification Algorithm (Decision Tree, Random Forest Naive Bayes,etc.) applied
Performance evaluation based on accuracy
Result/Prediction

Fig.1: Model Diagram

Based on the financial record data, major supervised algorithms are used to forecast the behavior of customer i.e.,

- Decision Tree,
- Naive Bayes,
- Random forest,
- K-nearest Neighbors,
- Logistic Regression,
- Support Vector Machine.

CAP Curve

CAP stands for Cumulative Accuracy Profile. Larger the area covered between the random model (aR) and perfect model (aP) line is the better model than the other models. 'Cumulative Accuracy Profile' is used in the performance evaluation of the classification model.

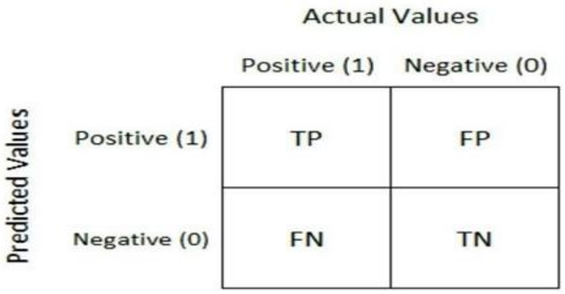
CLASSIFICATION REPORT

Classification report is most important parameter to evaluate the quality of predictions. It helps in measuring that how many predictions are True and how many are False. Various parameters of classification report are:

- Precision is fraction of tp / (tp + fp).
  The precision defines correctness of the classification model.
- Recall is fraction of tp / (tp + fn).
  The recall represents comprehensiveness of model.

CONFUSION MATRIX

Confusion matrix is an important tool to evaluate the production of a model. The x –axis of the matrix depicts instances in a predicted class and the y-axis shows the instances in an actual class.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig.2 Confusion Matrix

Some related terminologies we have to understand such as TP, FP. TN and FN in terms of fraudulency:

1. True Positive:

You can easily forecast that a customer is double-dealing, and he/she absolutely is.

2. True Negative:

You can easily conclude in case a customer is not scam, he/she veritably is denial.

3. False Positive:

You can easily forecast whether a customer is fraud, but he/she really is refused.

4. False Negative:

You can figure out that a customer is not fake, but he/she truly is.

Accuracy Score

Accuracy score is the defined as the fraction of predictions, our model got right. If there is a high accuracy score, better the model will be.

Accuracy = tp+tn/tp+tn+fp+fn

3.2 DEFINING THE DATA SET

- credit.policy : If client fulfills the credit, returns 1 otherwise 0.
- purpose: The objective of the loan is cited.
- int.rate: The lending rate is referred in this. Borrowers judged to be insecure having high price of money.
- installment: If loan is funded, the installments are owned by the borrower.

- log.annual.inc: Details of the yearly revenue of the debtor.
- dti: The debt-to-income ratio is cited.
- fico: It defines the credit score of the debter.
- days.with.cr.line : The number of days the debtor has had a credit line.
- revol.bal: In end of the credit card process, the amount of borrower that is unpaid is referred.
- revol.util : In this, the amount of the credit line used relative to total credit available is mentioned.
- inq.last.6mths : Exploration of debtor in last six thirty days.
- delinq.2yrs: The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
- pub.rec: Public records are listed.
- not.fully.paid : It returns 1 means borrower is not going to pay the loan completely otherwise 0 that is very important to check all the amount paid or not.

TRAINING AND TESTING OF THE DATA

Decision Tree: Decision trees are schematical way of all the feasible solutions of a decision-based problem. It is used to make smart decision. First task is to design the example of DecisionTreeClassifier() and fit it to the training data and then form predictions from the test set and evaluate the confusion matrix and accuracy score.

Naïve Bayes: Naïve Bayes is based on the Bayes Theorem for calculating probabilities and conditional probabilities. A Naïve Bayes classifier assumes that the presence of particular feature in a class is unrelated to the presence of any other feature. First task to design an example of NaïveBayesClassifier() and fit it to the training data and then form predictions from the test set an evaluate the confusion matrix , accuracy score.

K-Nearest Neighbour (KNN): It is the machine learning algorithm in which classification of object takes place by a majority rate of its neighbours. In this learning model, first of all design an example of K-NearestNeighbourClassifier() and fit it to the training data and then form predictions from the test set and evaluate the confusion matrix , accuracy score.

Random Forest: It is the most popular model in terms of accuracy. In this learning model, first task is to design an example of the RandomForestClassifier() and fit it to the training data and then form predictions from the test set and evaluate the confusion matrix and accuracy score.

Logistic Regression: It is a predictive analysis algorithm and based on probability. First task is to design an example of LogisticRegressionClassifier() and fit it to the training data and then form predictions from the test set and evaluate the confusion matrix, accuracy score.

Support Vector Machine (SVM): In this algorithm, analysis of data for classification and regression. This algorithm outputs a map of sorted data with the margins between the two as far apart as possible. First step is to design an example of SupportVectorClassifier() and fit it to the training data and then form predictions from the test set and evaluate the confusion matrix ,accuracy score.

3.3 CALCULATIONS:

1. When test_size=0.30

Classification Application	Accuracy Score	F-Score		False-positive	False-negative
		0	1		
Decision Tree	72.89%	0.84	0.21	407	372
Naïve Bayes	82.28%	0.90	0.15	76	433
Random Forest	83.33%	0.91	0.03	9	470
k-nearest Neighbor	81.35%	0.90	0.06	75	461
Logistic Regression	83.47%	0.90	0.06	2	473
Support vector Machine	83.40%	0.91	0	0	477

In this, when test_size =0.30 and train_set=0.70 we calculated the accuracy score,f-score,false-positive and false-negative.In logistic regression accuracy score is very high but SVM has f-score value zero so the perfect model is Support vector machine in this case.

2. When test_size=0.25

Classification Application	Accuracy Score	F-Score		False-positive	False-negative
		0	1		
Decision Tree	72.47%	0.85	0.21	326	283
Naïve Bayes	82.11%	0.90	0.13	97	335
Random Forest	84.69%	0.92	0.04	9	359

k-nearest Neighbor	82.15%	0.90	0.07	81	350
Logistic Regression	84.62%	0.92	0.04	7	358
Support vector Machine	84.59%	0.92	0	0	366

When test_size =0.25 and train_set=0.75 by calculating all the important parameters to determine which model is best. The accuracy score of the logistic regression is very high (84.62%) but F-score of the support vector machine has a value zero.So,in this case the best learning model is Support vector machine.

3. When test_size=0.20

Classification Application	Accuracy Score	F-Score		False-positive	False-negative
		0	1		
Decision Tree	73.64%	0.84	0.21	281	224
Naïve Bayes	81.88%	0.90	0.13	80	267
Random Forest	84.65%	0.92	0.03	6	288
k-nearest Neighbor	81.99%	0.90	0.07	65	280
Logistic Regression	84.70%	0.92	0.04	6	287
Support vector Machine	84.70%	0.92	0	0	293

When test_size =0.20 and train_set =0.80 by calculating all the parameters we came to conclusion that accuracy score of both learning models (logistic regression and Support vector Machine) is same but value of f-score of logistic regression is 0.04 and SVM is 0. So the perfect model is support vector machine in this case too.

4. When test_size=0.15

Classification Application	Accuracy Score	F-Score		False-positive	False-negative
		0	1		
Decision Tree	74.53%	0.85	0.22	206	163
Naïve Bayes	82.25%	0.90	0.14	59	196
Random Forest	84.82%	0.92	0.04	2	212
k-nearest Neighbor	82.11%	0.90	0.07	51	206
Logistic Regression	85.03%	0.92	0.05	5	210
Support vector Machine	84.96%	0.92	0	0	216

When test_size =0.15 and train_set=0.85 by calculating all the parameters we came to that point accuracy score of logistic regression is high (85.03%) but the F-score of the SVM has a value zero. So, the best learning model is Support vector machine.

5. When test_size=0.10

Classification Application	Accuracy Score	F-Score		False-positive	False-negative
		0	1		
Decision Tree	73.27%	0.84	0.18	150	106
Naïve Bayes	82.88%	0.91	0.13	42	122
Random Forest	86.32%	0.93	0.07	2	129
k-nearest Neighbor	83.19%	0.91	0.07	33	128
Logistic Regression	86.01%	0.92	0.06	4	130
Support vector Machine	85.51%	0.92	0	0	134

When test_size =0.10 and train_size =0.90, by calculating the important parameters we conclude that accuracy score of the logistic regression is very high (86.01%) but the F-score of support vector machine has a value zero, so the best recommended learning model is Support Vector Machine.

3.4 CAP curve analysis:

With the help of CAP curve analysis, we can easily justify the quality of our models. The ratio lies between 0 and 1. More the ratio is closer to one, best the learning model is by calculating the accuracy score of all the learning models, the accuracy score of the logistic regression is highest (83.47%) as compared to all the learning models but the F-score of the Support vector machine has a value zero.

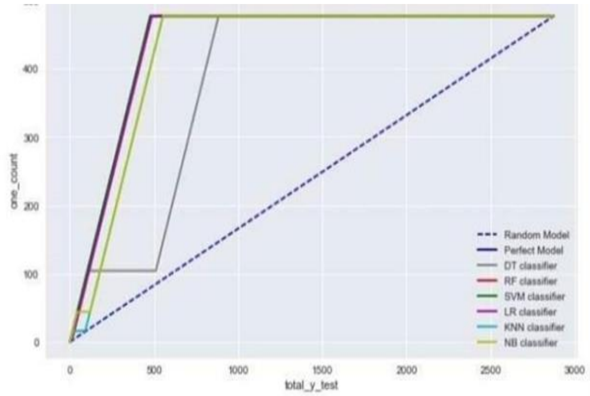


fig3.CAP Curve analysis of all learning models

4. RESULT AND DISCUSSION

The accuracy score of the different classifiers that we have got are 72.89%, 82.28%, 83.33%, 81.35%, 83.47% and 83.40% for Decision Tree, Naive Bayes, Random Forest, K-nearest Neighbors, Logistic Regression and Support Vector Machine Classifiers, respectively. Here, we can conclude that Logistic Regression is better. But accuracy is not only factor to consider the best model because in our case, but we also have to consider false-positive in confusion matrix and try to minimize it. In this case, Support vector machine is better among all the models. Also, Support Vector Machine has higher f-score as compared to Logistic Regression. After analyzing the CAP curve, we have seen that the line of SVM model overlaps the perfect model which means that it is better classifier than the other classifiers. Finally, we conclude that Support Vector Machine performed better than the other classifiers.

5. CONCLUSION

Based on the accuracy score, f-score, CAP curve & false-positive parameter of this project, we came to that point, Support Vector Machine model performed good than the other models. Banks can use this project model into their system to predict that whether the customer is fraud or not. By doing this, banks can have more chance to save themselves from any scams or fraudulency in future.

REFERENCES

[1] X.Francis Jency, V.P. Sumathi, Janani Shiva Sri, An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clientsʿ IJRTE, Volume-7, Issue-4s, Nov 2018.

[2] Shweta Yadav, Sanjeev Thakur, ‚Bank Loan Analysis using Customer Usage Data: A Big DATA Approach Using Hadoopʿ, IEEE, 2nd International Conference on Tel-NET, 2017.

[3] Y. Sachin, E. Duman, Detecting Fraud by Decision Tree and Support Vector Machineʿ, In Proceedings of the international multi-Conference of Engineers and Computer Scientists, Hong Kong, 2011, pp.1-6.

[4] D.S. Sisodia, N.K.Reddy, and S.Bhandari, "Performance Evaluation of class Balancing Techniques for Fraud detection," IEEE Int. Conf. Power, control. Signals Instrum Eng, pp2747-2752, 2017.

[5] J. Han, M.Kamber, ‚Data Mining: Concepts and Techniquesʿ, Seconded, Morgan Kaufmann Publishers, 2006, pp. 285–464.

[6] L.Breiman. Random forests. Machine learning, 45(1):5–32, 2020.

[7] A. Bahnsen, D. Aouada, A. Stojanovic and B. Ottersten, “Feature Engineering Strategies for Credit

[8] Card Fraud Detection", ELSEVIER Expert System with Applications, pp. 134-142, 2016.

[9] Y. Sachin and E. Duman, "Detecting Credit Card Fraud by Decision Tree and Support Vector Machine", Proceedings of the international multi-Conference of Engineers and Computer Scientists, pp. 1-6, 2011.

[10] E. Ngai, Y. Hu, Y. Wong, Y. Chen and X. Sun, "The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature", Elsevier Decision Support Systems, pp. 559-569, 2011.