Effective Credit Default Scoring using Anomaly Detection

Krunal M. Surti
PG Student
Department of Computer Engineering
Silver Oak College of Engineering & Technology

Mr. Ashish Patel
Assistant Professor
Department of Computer Engineering
Silver Oak College of Engineering & Technology

Abstract

In recent years there has been a trend towards online purchase so stealing of credit data is high like identity of the credit card owner, password or etc. the attacker may use this data for to take loan from financial domain and they make credit default. Credit scoring is the give the creditworthiness of person. Anomaly Detection is the process of classifies unusual behavior. It is important data analysis task used for classify interesting and emerging patterns, trends and anomalies from data. Anomaly detection is an important tool to detect irregularity in many different domains including financial fraud detection, computer network intrusion, human behavioral analysis and many more. In today's era the credit and Loan Default is become high because of many fraudulent activity or increase online purchases. To perform anomaly detection in this paper linear regression with rule based classification and logistic regression is used. The preprocessing is used for to perform explore, analyze and determine the factor that play crucial role to find credit default.

Keywords: Anomaly Detection, Credit Default, Credit Score, Creditworthiness, Preprocessing, Regression

I. Introduction

Anomaly Detection is the classifying of things, events or observations that don't change to normal behavior. Those autonomous arrangements are usually mentioned as anomalies, outliers in numerous domains. Anomaly detection reveals countless use in an exceedingly extensive variety of uses like fraud detection for credit cards, insurance, or health care, intrusion detection for cyber-security against crime, fault detection in protection important systems, associated army police investigation for enemy activities the instance of anomaly detection is an computer community network that comprise abnormal website guests may suggest that a hacked system is inflicting out sensitive data to associate unauthorized destination. An anomalous MRI photograph might also display the closeness of threatening cancer tumors. Inconsistency in credit card transaction data may want to illustrate credit card or identification theft, or a satellite space craft sensing information comprise abnormal data may signify a fault in some part of the space craft.

Anomalies are integrate in statistics that don't conform to a nicely delineate approach of normal behavior. Fig. 1 suggests that the anomaly is an exceedingly straightforward 2-dimensional data set. The data has 2 traditional clusters, X and Y, since most observations lie in these two regions. Points that are sufficiently far away from the regions, e.g., point's A and B and points in region C are anomalies. [12]

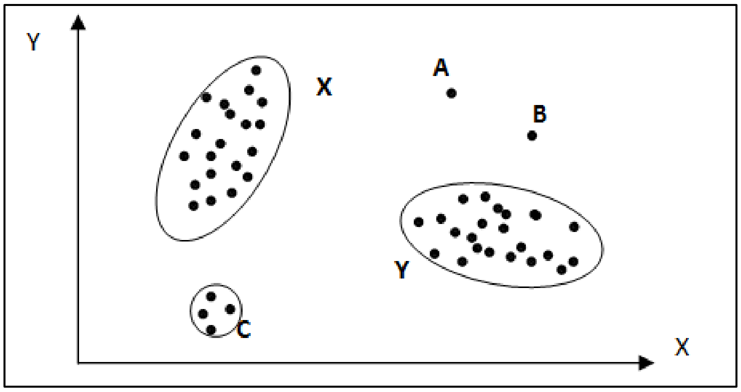

Fig. 1: Point Anomaly [12]

A credit score may be a statistical number that depicts a person's trustiness. Lenders use a credit score rating to assess the probability that a person repays his owed. All the organizations will generate a credit score rating for every person by victimisation the person's preceding credit history. A credit score may be a 3-digit number range begins from 300 to 850, with 850 is the highest score that a borrower achieves. If the score of person is higher than the more financially trustworthy a person is considered to be. In simple language, a credit/loan default is given by a borrower have not made their agreed upon credit/loan payments to the lender. There are many varieties of reasons why a client may not have made payments, however a given period of time has passed, that non-payment record will become a part of the consumer's credit history. After it becomes a part of the borrower's credit history (or credit record) it is used during the formulation of the consumer's credit score.

The majorities of tasks associated with credit fraud are created within the dark web websites and specialized to s different services in the dark web. These environments allow the flooding of illegal activities that related to the commercialization of stolen credit and debit cards and related information. The dark web represents a part of internet that is considered vital for the business of criminal crews that specialize in performing different fraudulent activities related to credit cards.In the dark web different communities provide numerous things and illegal services, such as many amount of stolen card data, the codes that used for adjustment of payment systems (i.e. PoS, ATM), and card on demand services. In these dark web markets the offender will effortlessly gather and sell tools, illegal services and dataset that use for different types of illegal activities. The bank accounts opened with fake identities this fake account is then used for payment recipients for the sale of any type of product and service that related to credit card fraud. [13]

II. Related work

In [1] the Author introduced Model for finding Anomaly in network using k-means clustering machine based approach with the use of big data analytical techniques and other approach is to find the best results to prevent attacks at it's very origin, in this paper they used a HDFS in spark shell they used R tools for visualization. In paper [2] the technique uses graph-based descriptors to capture the network billing activity for a specific time period, where nodes in the graph represent users and/or servers, and edges represent communication events. This paper presents the method followed to create multiple graphs in each vertex neighbourhood of predefined size and presents the features that are extracted from the neighbourhood graphs for different neighbourhood sizes k, and used in order to train a supervised classification algorithm, namely, the Random Forest (RF) classifier, to recognize anomalous graphs. In paper [3] describes a transforming model that converts every entry into vector. Every value in the vector is a probability value, that is, every feature of each attribute is transformed to a corresponding value by statistical techniques or Naive Bayes. They propose a method to deal with URI without any query. It splits URI path string into tokens, so applies Naive Bayes to get their probability value. The detection system is carried out by a real-life dataset of millions of entries which is much larger than the datasets of prior study.

In paper [4] gives the comparison of three algorithms given by scaled conjugate gradient back propagation, Levenberg-Marquardt and One-step secant back propagation (SCG, LM and OSS). Different parameters have been used for experiment to try and do the comparison; training time, gradient, MSE and R. The slow algorithm is OSS; the algorithm with the biggest gradient is SCG. The best algorithm is LM because it has the biggest R, this is best for this dataset. This paper [5] defines the feature selection method and also uses random forest method and gives attention to analysis propose the selected social and economic factors additionally to the bank typically used ones. This may help creditors understand the determinants of default risk. This Paper analyzes a massive dataset from the Lending Club. The data were collected from 2007 to 2011, this data include the effect of economic crisis in 2008 In this paper three kinds of DM models (DT, NN and SVM) and also use 10-fold cross validation and two classification metrics to evaluate the prediction results. The Paper [6] exploratory examine is meant to deal the problem of fraudulent loan requests on peer-to-peer (P2P) platforms. They suggest a collection of alternatives that capture the behavioral characteristics (e.g., learning, past performance, social networking, and herding manipulation) of malevolent borrowers, who intentionally create loan requests to accumulate finances from lenders but default later on. They determined that using the widely adopted classification methods such as Random Forest and Support Vector Machines, the proposed feature set outperform the baseline feature set in helping detect fraudulent loan requests.

The paper [7] gives the approach of Logistic Regression also use Classification and Regression Trees (CART) with techniques including under sampling, Prior Probabilities, Loss Matrix and Matrix Weighing to address unbalanced records. The paper [8] aims to develop a model and construct a prototype for the same using a data set available in the UCI repository. The model is a decision tree based classification model that uses the functions available in the R Package. Before to building the model, the dataset is pre-processed, reduced and made ready to provide efficient predictions. The final model is used for prediction with the test dataset and the experimental results prove the efficiency of the built model. The paper [9] Ensemble machine learning algorithms and pre-processing techniques are used to examine, analyze and conform the factors that play crucial role to predict the credit risk concerned in "Lending Club" 2013- 2015 loan applications dataset. A loan is examined "good" if it's repaid with interest and on time. The algorithms are optimized to favor the potential good loans whilst identifying defaults or risky credits. The paper [10] suggest a model to evaluate the finding risk of default, based on the analysis, they could appropriately assess the risk of default enhance return on investment.

III. Methodology

Loan default is become major issue now a days this paper present the technique for predict the loan default and also give technique to find loan defaulter in future, the dataset is used here is the lending club for the data contains the 2 lakh records and multiple attributes. The data preproceessing is on dataset to perform the data cleaning or remove some attribute that are not useful. The imputation technique is perform to impute missing value in dataset after that the random forest is used select the variables after the random forest execute the correlation is used to give the final attribute, in actual data set one more column namely Monthly_inc is added. The all the charts and table is given below, the R programming is used to find the loan default.

A. Data Cleaning

There are 27 variables in loan data, First I did a data cleaning job in loan application almost there are all the column is complete but there one column that is in complete so I used imputing method to replace missing value, There is package called MICE (Multivariate Imputation By Chained Equations) in this package there is method called Predictive Mean Matching for imputation of missing value, the loan database is given with loan information and personal information of loan

Table – 1
Before imputation

NO.	*Mths_since_last_delinq*	*dti*
1	*42*	*14.92*
2	*NA*	*12.02*
3	*60*	*18.49*
4	*NA*	*25.81*
5	*17*	*8.31*
6	*NA*	*39.79*

Table – 2
After imputation using MICE package

NO.	*Mths_since_last_delinq*	*dti*
1	*42*	*14.92*
2	*77*	*12.02*
3	*60*	*18.49*
4	*38*	*25.81*
5	*17*	*8.31*
6	*19*	*39.79*

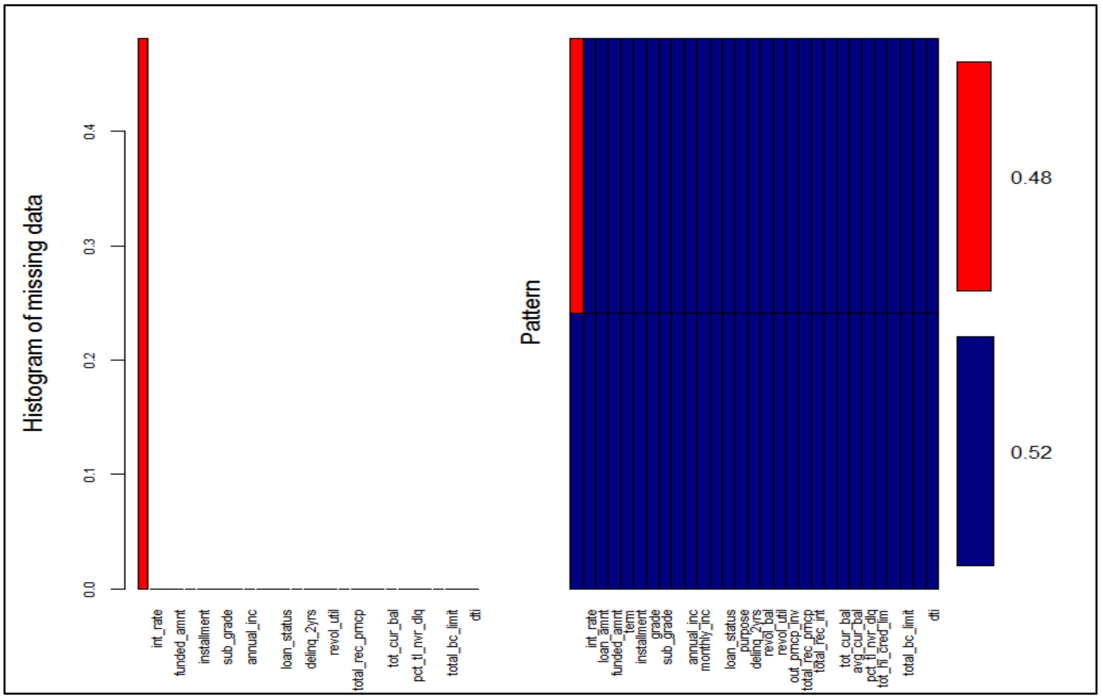

Fig. 2: Missing Data graph By VIM Package

IV. FEATURE SELECTION

In feature selection there is different method used but here I used the Random Forest and the Correlation method for feature selection. In R there is Boruta package that used for building a Random Forest tree for feature selection at the last it give the box plot graph of all the input variables if variable is correct than Boruta package give Confirmed statement if not it give rejected statement, after random forest selection I also used correlation function to perform feature selection the last selected variables is give below table.

Table - 3
Selected variable using Random forest and correlation

Variable	*Attribute*
int_rate	*Interest rate on the loan*
term	*The number of payments on the loan. Values are in months 36 or 60.*
grade	*Loan grade*
sub grade	*Loan subgrade*
revol_util	*The amount of credit the borrower is using relative to all available revolving credit.*
revol_bal	*Total credit revolving balance*
loan_amnt	*The amount of the loan taken by the borrower*
Installment	*The monthly payment owed by the borrower if the loan originates.*
annual_inc	*The annual income of loan borrower*
last_pymnt_amnt	*Last total payment amount received*
total_rec_int	*Interest received to date*
total_rec_prncp	*Principal received to date*
last_pymnt_amnt	*Last payment amount received*

V. RESULTS AND DISCUSSIONS

The liner regression graph for anomaly detection is given below, the linear regression graph is for monthly_income vs. installment now here I give the regression graph of the 200 value for simplicity. Now the anomaly is detected using by simple probability. If the monthly income of borrower is less but it take the highest loan than it likely to be perform default. The given dataset is contain 2 lakh records so first I divide the dataset in to 1000 records of data in this paper I give the detected anomaly of first 5 part of data. The simple algorithm for find anomaly is given below.

A. *Steps for Find Anomaly*

- Step 1 – Select set of n records
- Step 2 – create Subset s1 where charged off/default = TRUE
- Step 3 – create subset s2 where charged off/default =FALSE
- Step 4 – calculate mean of s1 as Ms1
  - MS1=mean (s1)
- Step 5 – calculate mean of s2 as Ms2
  - Ms2=mean (s2)
- Step 6 – find difference of Ms1 and Ms2
  - D = Diff (Ms1, Ms2)
- Step 7 – calculate Ms from original data
  - Ms=mean (Data)
- Step 8 – X = ifelse (Ms > D, Ms, D)
- Step 9 – for find Y repeat step 5 to 8
- Step 10 – Y = ifelse (Ms > E, Ms, E)
- Step 11– ifelse (installment > X, loan_status == Charged Off/Default)
  - ifelse (monthly_inc > X, loan_status == Charged Off/Default)
  - ifelse (monthly_inc > X, installment < Y, loan_status == Charged Off/Default)

Rule 1: DF <- ifelse(data$installment >= X & data$loan_status == "Charged Off", " Charged off", "not charged off")

In above rule the value X is used now as the given rule if installment is greater than or equal to/less than or equal to X and the loan status is “Charged Off “ than it give the anomaly in loan data.

Table – 4
Classified Value Table for Installment Vs. Loan_status

Data with (1000 Rows)	*Mean Value (Ms1)*	*Mean Value (Ms2)*	*Ms*	*Actual*	*Classified*
Part1	*1207.32*	*414.50*	*466.08*	*65*	*64*
Part2	*1006.77*	*447.29*	*492.05*	*80*	*80*
Part3	*896.62*	*426.02*	*451.73*	*58*	*56*
Part4	*799.61*	*450.20*	*421.94*	*91*	*86*
Part5	*732.83*	*421.94*	*448.03*	*84*	*83*

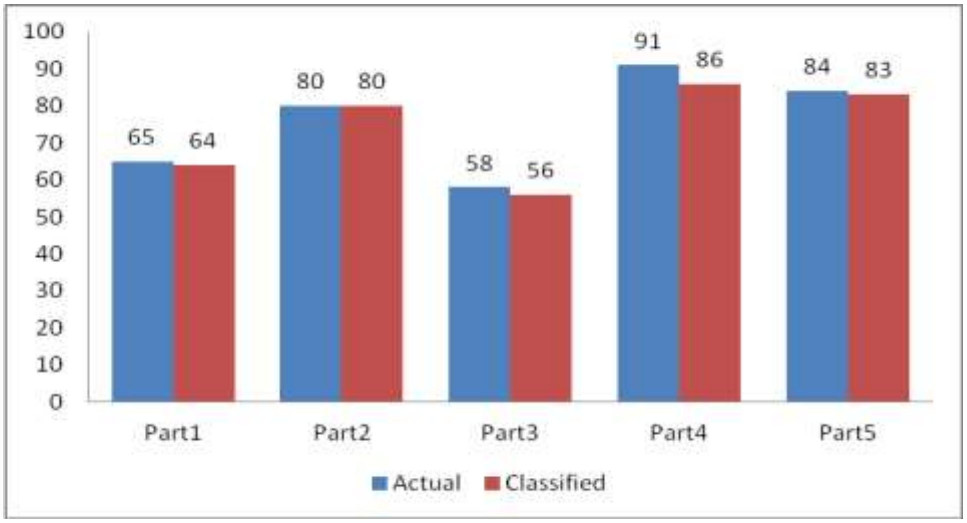

Fig. 3: Comparison Chart of Installment Vs. Loan_status for Actual And Classified Value.

Table – 5
Classification Table for Installment Vs. Loan _status

Rows	*Classification*
1	*Not Charged Off*
2	*Not Charged Off*
3	*Not Charged Off*
4	*Not Charged Off*
5	*Charged Off*

Rule2: DF <-ifelse(data$monthly_inc <= Y & data$installment >= X & data$loan_status == "Charged Off", "Charged off", "not charged off")

In above rule the value of X and Y used as the given rule if installment is greater than or equal to X and the monthly-inc is less than the Y and loan status is “Charged Off “ than it give the anomaly in loan data

Table – 6
Classified Value Table for Monthly_inc vs. Installment vs. Loan_status

Data with (1000Rows)	Actual	Classified
Part1	65	12
Part2	80	28
Part3	58	14
Part4	91	41
Part5	84	43

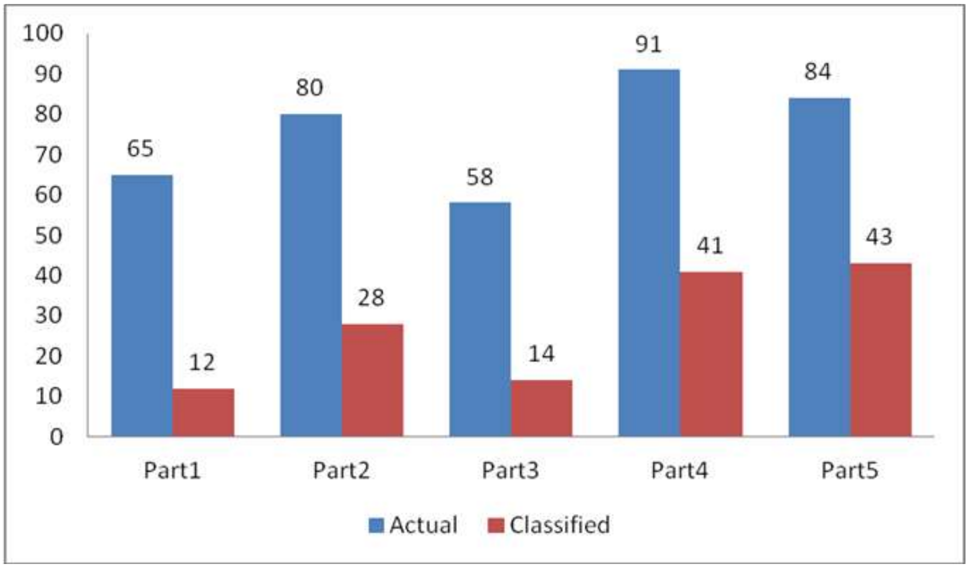

Fig. 4: Comparison Chart of monthly_inc Vs. Installment Vs. Loan_status for Actual And Classified Value

Table – 7
Classification Table monthly_inc vs. Installment vs. Loan_stauts

Rows	*Classification*
1	*Not Charged Off*
2	*Not Charged Off*
3	*Not Charged Off*
4	*Not Charged Off*
5	*Not Charged Off*

Rule3: DF <-ifelse(data$monthly_inc >= X & data$loan_status == "Charged Off", "Charged off", "not charged off")In above rule the value Y is used now as the given rule if monthly_inc is greater than or equal to/less than or equal to Y and the loan status is “Charged Off “than it give the anomaly in loan data.

Table – 8
Classified Value Table for monthly_inc Vs. Loan_status

Data with (1000 Rows)	*Mean Value (Ms1)*	*Mean Value (Ms2)*	*Ms*	*Actual*	*Classified*
Part1	*10655.79*	*6530.94*	*6799.32*	*65*	*52*
Part2	*10173.88*	*7063.46*	*7312.30*	*80*	*52*
Part3	*9110.76*	*6355.46*	*6515.73*	*58*	*42*
Part4	*8660.27*	*6646.66*	*6829.90*	*91*	*45*
Part5	*7229.26*	*6448.24*	*6513.85*	*84*	*43*

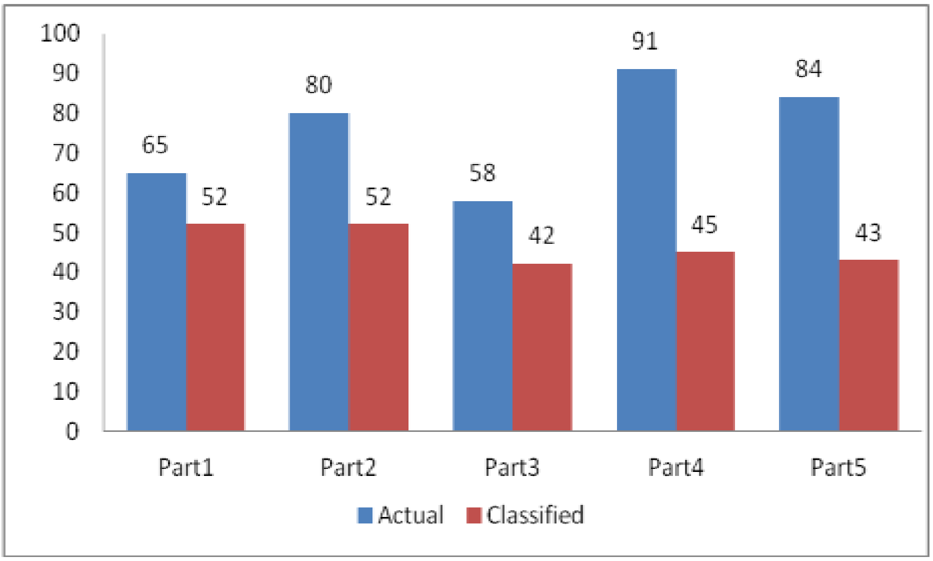

Fig. 5: Comparison Chart of monthly_inc Vs. Loan_status for Actual and Classified Value

Table – 9
Classification Table for monthly_inc Vs. Loan_status

Rows	*Classification*
1	*Not Charged Off*
2	*Not Charged Off*
3	*Not Charged Off*
4	*Not Charged Off*
5	*Not Charged Off*

B. Logistic Regression

The logistic regression is used to estimate the probability of binary response based on one or more predictor variables. Probability equations of logistic regression is

$$\text{Probability} = 1/(1 + e\hat{}((coefficient * X + intercept)))$$

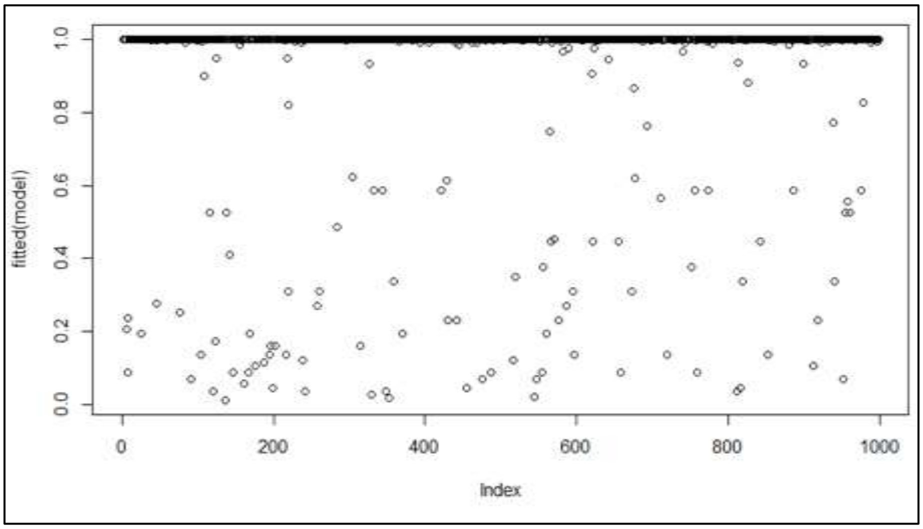

Fig. 6: Fitted Value Chart for Logistic regression

Now Find Fitted value for loan data the below given tables is give the fitted value for loan_status Vs. Installment.

Table – 10
Comparison Chart for Actual and Classified Value Loan_status Vs. Installment

Data with (1000 Rows)	*Actual*	*Classified*
Part1	*65*	*69*
Part2	*80*	*61*
Part3	*58*	*20*
Part4	*91*	*25*
Part5	*84*	*17*

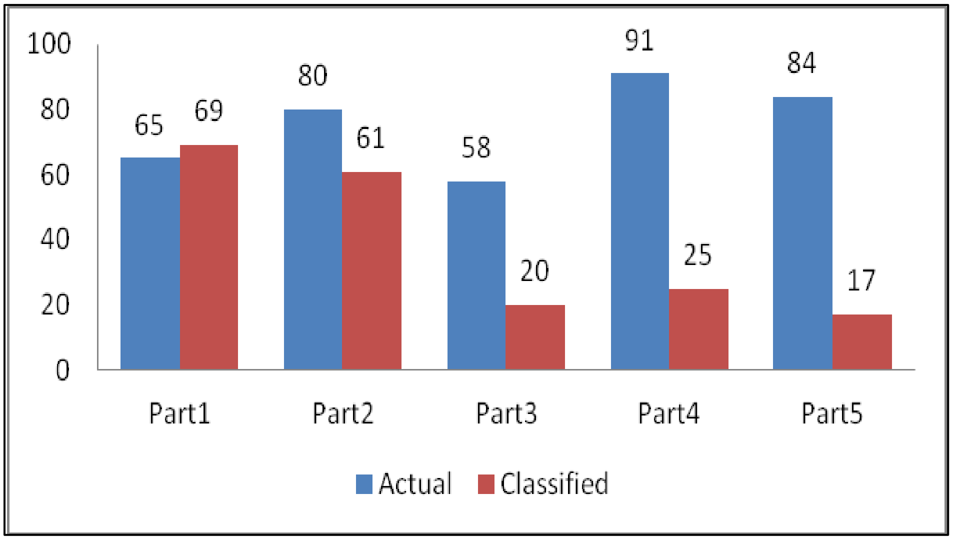

Fig. 7: Comparison Chart for Actual and Classified Value For Logistic regression

Table – 11
Comparison Chart for Actual and Classified Value Loan_status Vs. Monthly_inc

Data with (1000 Rows)	*Actual*	*Classified*
Part1	*65*	*15*
Part2	*80*	*3*
Part3	*58*	*2*
Part4	*91*	*1*
Part5	*84*	*0*

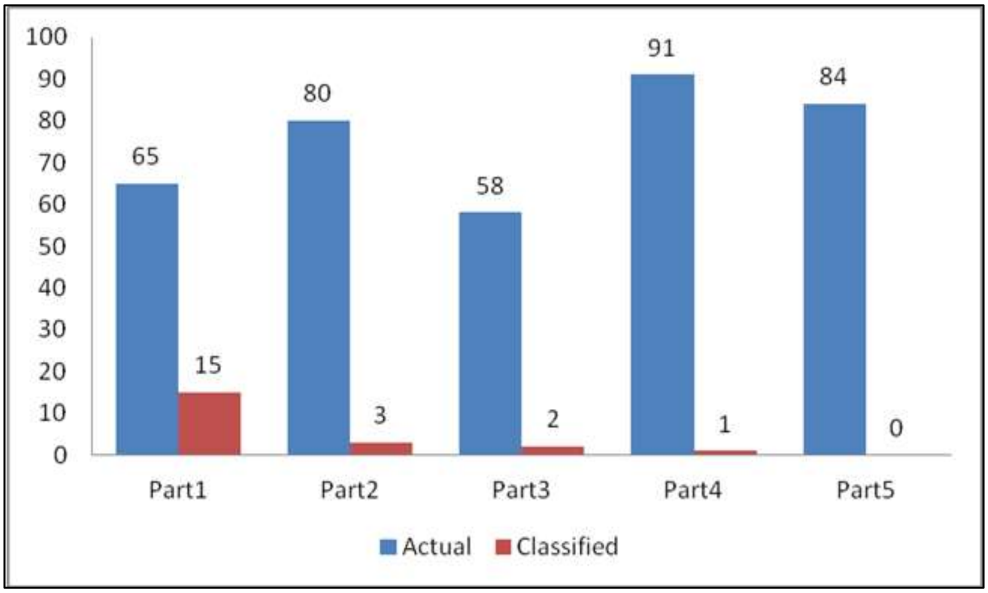

Fig. 8: Comparison Chart for Actual and Classified Value for Logistic regression

Table – 12
Comparison Chart for Actual And Classified Value Loan_status Vs. Monthly_inc Vs. Installment

Data with (1000 Rows)	*Actual*	*Classified*
Part1	*65*	*73*
Part2	*80*	*62*
Part3	*58*	*17*
Part4	*91*	*13*
Part5	*84*	*14*

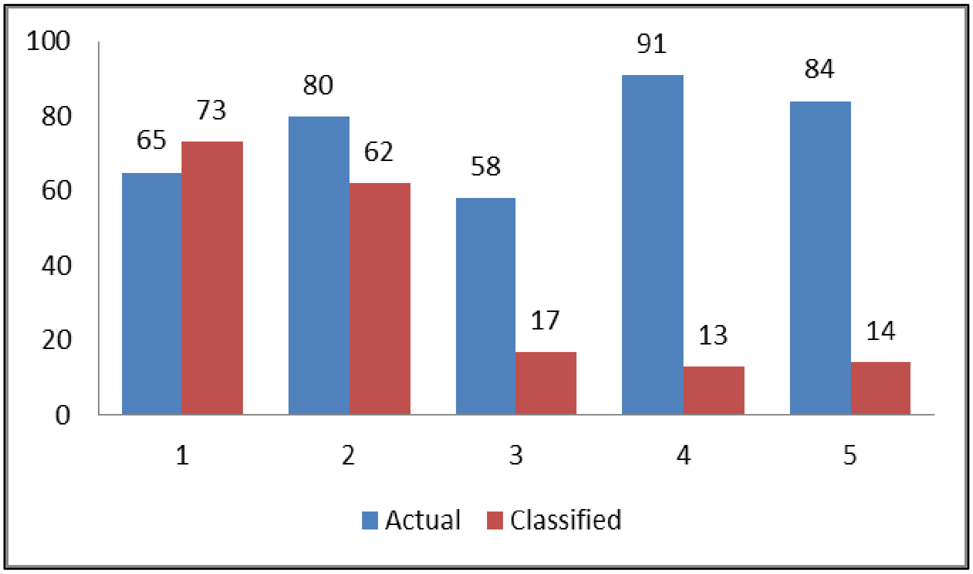

Fig. 9: Comparison Chart for Actual and Classified Value for logistic regression

C. *Comparison Chart*

The Given Below figure shows that rule based classifier is more accurate as compare to logistic regression.

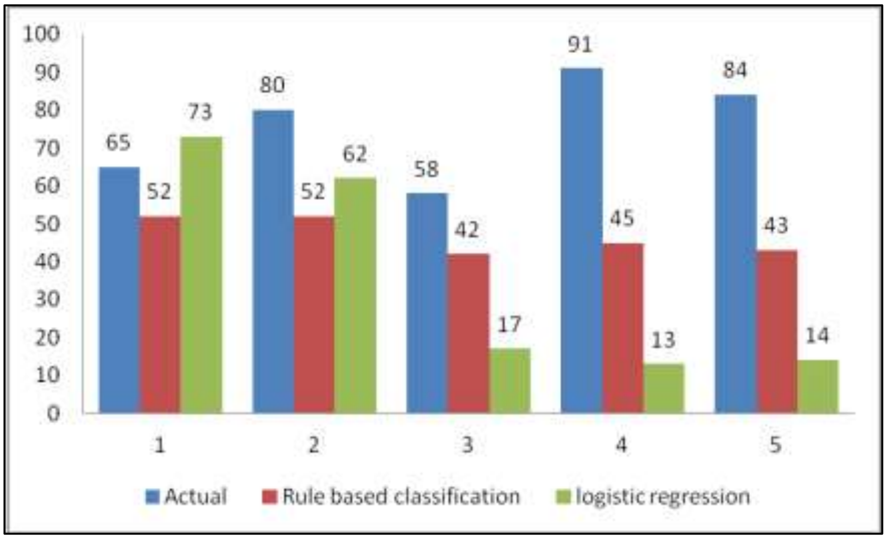

Fig. 10: Comparison Chart for Logistic regression vs. Rule based classification

The Given Below figure shows that rule based classifier is more accurate as compare to logistic regression.

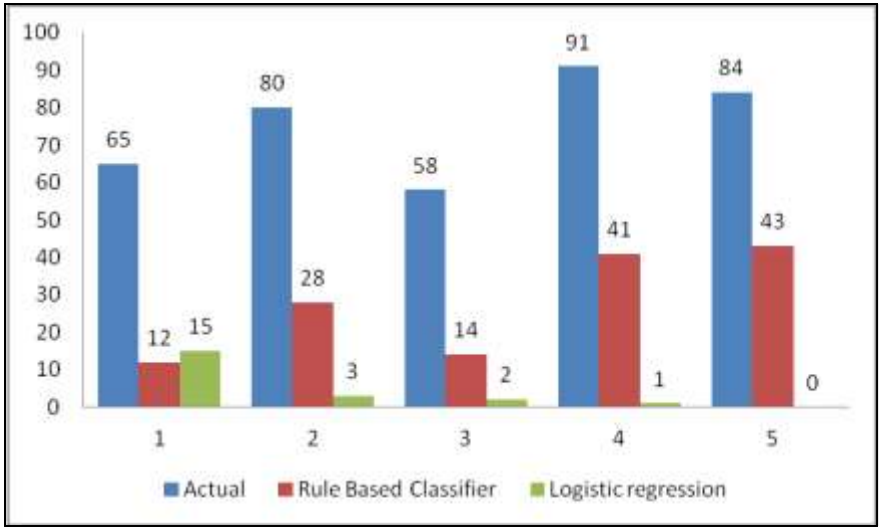

Fig. 11: Comparison Chart for Logistic regression vs. Rule based classification

The Given Below figure shows that rule based classifier is more accurate as compare to logistic regression

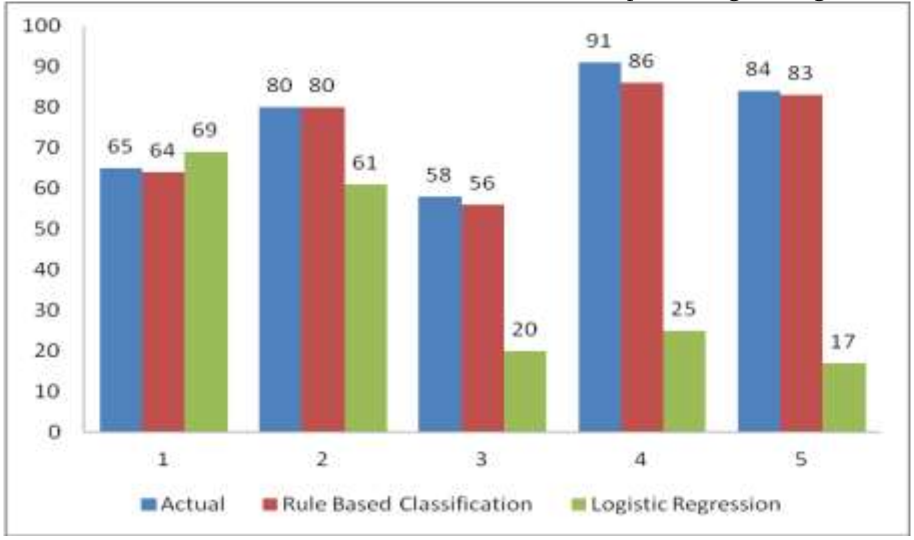

Fig. 12: Comparison Chart for Logistic regression vs. Rule based classification

So as compare to logistic regression the Rule based give more accurate classification of the Loan defaulter, using this rule user can also find future prediction that the borrower is likely to be performing default.

VI. Conclusion

In this paper I briefly review about the anomaly detection and the different techniques, For anomaly detection I give the rule based classification to perform the classification, for perform classification the algorithm is made using the three rule the all rule give the classified data from the loan the comparison chart also given, the logistic regression is also perform for comparing classified data as per comparison with logistic regression and rule based classification the rule based classification is give much better result as compared to logistic regression.

References

[1] R. Kumari, Sheetanshu, M. K. Singh, R. Jha and N. K. Singh, "Anomaly detection in network traffic using K-mean clustering," 2016 3rd International Conference on Recent Advances in Information Technology (RAIT), Dhanbad, 2016, pp. 387-393. doi: 10.1109/RAIT.2016.7507933.

[2] S. Papadopoulos, A. Drosou and D. Tzovaras, "A Novel Graph-Based Descriptor for the Detection of Billing-Related Anomalies in Cellular Mobile Networks," in IEEE Transactions on Mobile Computing, vol. 15, no. 11, pp. 2655-2668, Nov. 1 2016.doi: 10.1109/TMC.2016.2518668

[3] S. Zhang, B. Li, J. Li, M. Zhang and Y. Chen, "A Novel Anomaly Detection Approach for Mitigating Web-Based Attacks Against Clouds," 2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing, New York, NY, 2015, pp. 289-294. doi: 10.1109/CSCloud.2015.46

[4] K. I. Hassan and A. Abraham, "Modeling consumer loan default prediction using neural netware," 2013 INTERNATIONAL CONFERENCE ON COMPUTING, ELECTRICAL AND ELECTRONIC ENGINEERING (ICCEEE), Khartoum, 2013, pp. 239-243. doi: 10.1109/ICCEEE.2013.6633940

[5] Y. Jin and Y. Zhu, "A Data-Driven Approach to Predict Default Risk of Loan for Online Peer-to-Peer (P2P) Lending," 2015 Fifth International Conference on Communication Systems and Network Technologies, Gwalior, 2015, pp. 609-613. doi: 10.1109/CSNT.2015.25

[6] J. Xu, D. Chen and M. Chau, "Identifying features for detecting fraudulent loan requests on P2P platforms," 2016 IEEE Conference on Intelligence and Security Informatics (ISI), Tucson, AZ, 2016, pp. 79-84. doi: 10.1109/ISI.2016.7745447

[7] S. Birla, K. Kohli and A. Dutta, "Machine Learning on imbalanced data in Credit Risk," 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, 2016, pp. 1-6. doi: 10.1109/IEMCON.2016.7746326

[8] G. Sudhamathy and C. J. Venkateswaran, "Analytics using R for predicting credit defaulters," 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, 2016, pp. 66-71. doi: 10.1109/ICACA.2016.7887925.

[9] Vinod Kumar L, Natarajan S, Keerthana S, Chinmayi K M and Lakshmi N, "Credit Risk Analysis in Peer-to-Peer Lending System," 2016 IEEE International Conference on Knowledge Engineering and Applications (ICKEA), Singapore, 2016, pp. 193-196. doi: 10.1109/ICKEA.2016.7803017

[10] Lei Xia and Jun-feng Li, “Analysis on Credit Risk Assessment of P2P” 2016 E. Qi et al. (eds.), Proceedings of the 22nd International Conference on Industrial Engineering and Engineering Management 2015, DOI 10.2991/978-94-6239-180-2_86.

[11] Chandola, V., Banerjee, A., and Kumar, V. 2009. Anomaly detection: A survey. ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 pages. DOI = 10.1145/1541880.1541882

[12] http://www.investopedia.com/terms/d/defaultrisk.asp.

[13] Jiawei Han, Micheline Kamber, Jian Pei, “Data Mining Concept and Techniques”.

[14] Sam Maes, Karl Tulys, Bram Vanschoenwinkel, Bernard Manderick, “credit card Fraud Detection. Applying Bayesian and Neural Network”, 2016.

[15] Ravinder Reddy, B.kavya, Y Ramadevi (Ph.D.), “A Survey on SVM Classifier for Intrusion Detection”, 2014.