A REPORT

ON

Credit Risk Management, Regression Analysis and Prediction of Credit Risk using Loan data

BY

Kapil Agrawal **2014B3A3579P** **B.E Electrical, MSc. Economics**

Prepared in partial fulfillment of the
Practice School-I Course

AT

Indian Bank, Chennai

A PRACTICE SCHOOL –I STATION OF

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI
(JUNE 2016)

A REPORT

ON

Credit Risk Management, Regression Analysis and Prediction of Credit Risk using Loan data

BY

Kapil Agrawal **2014B3A3579P** **B.E Electrical, MSc. Economics**

Prepared in partial fulfillment of the
Practice School-I Course

AT

Indian Bank, Chennai

A PRACTICE SCHOOL –I STATION OF

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI
(JUNE 2016)

BIRLA INSTITUE OF TECHNOLOGY AND SCIENCE, PILANI
Practice School Division

Station: Indian Bank
Centre: Chennai

Duration: 2 Months
Date Of Start: 23rd May, 2016
Date Of Submission: 16th July, 2016

ID No./Name: 2014B3A3579P / Kapil Agrawal
Discipline of the student: MSc. Economics / BE. Electrical Engineering

Name of the expert: Mr. Rakesh Dutt

Name Of the PS Faculty: Dr. Padma Murali

Keywords: Credit Risk, Regression, Prediction, Decision Trees, Confusion Matrix

Project Areas: Credit Risk Management, Regression, Machine Learning, Decision Trees

Abstract: In this world of finance, every financial institutions like banks need to take care of the risks that they are undergoing, especially credit risk. So, it's very useful to construct some credit risk model which aims at minimizing the credit risk and maximizing profits for the bank. Based on the previous big data of the customers and their loan details, its necessary to construct risk model which also predicts whether to give any customer a loan or not. This report also includes the regression analysis to analyze the relation between operational variables of the bank and an efficient way of analyzing stock prices of the stocks.

Signature of Student **Signature Of PS Faculty**

Date **Date**

ACKNOWLEDGEMENTS

I would like to thank the Assistant General Manager Mr. Satyendra Sharma and my guide Rakesh Sir of Risk management Department for sharing their valuable time and expertise to guide me throughout the project.

I would like to express my gratitude towards Indian Bank, for giving me such a wonderful opportunity to work with them and provided me the essential data for the analytics purpose.

I would also like to thank my PS instructor Dr. Padma Murali, Assistant Professor, Department of Mathematics for taking out time from her busy schedule to guide and supervise us in each and every aspect related to the PS-I programme.

I would like to thank the Practice School Division of our college for giving us this wonderful opportunity to apply our knowledge and skills in the outside world and learn a lot apart from our regular academic curriculum.

Finally, I would like to thank department members, friends and parents for their constant help and support.

TABLE OF CONTENTS (*Specimen*)

1. Organization Profile 6
2. Introduction 7
3. Scope and Objective 8

4. Credit Risk Management 9
5. Regression Analysis 10

6. Prediction of Credit Risk using Loan data 17
7. The Data 18
8. Approach required to analyze the data 20
9. Decision Trees 21
10. Chaid Analysis 23
11. Code of the project 24
12. Output of the code 26
13. Conclusions 30
14. References 31

ORGANISATION PROFILE

A bank is a place where it accepts deposits and channels them into lending activities. Banks become a financial intermediary through the activities they undertake. In general, the standard activities of a bank include:

- Conduct savings and current accounts for customer.
- Payment to customer against cheques and other negotiable instruments.
- Accepts term deposits.
- Issue debt securities like banknotes and bonds.

Provide loans for the purpose of housing, education, business, etc.

INDIAN BANK

Indian Bank is a nationalized bank in India. Indian Bank is now ISO 27001:2013 Certified. Indian Bank's Information Security processes have been successfully assessed for ISO 27001:2013 security standard and is among the very few Banks who are certified worldwide. Apart from being a best practice, the certification is a useful tool to add credibility and reassure the Indian Bank's clients that the Bank's information security is of high quality. With the state of the art technology, Indian Bank provides all banking services under one roof with 24*7 internet banking facility, 2532 branches and 2953 ATMs spread across all over India and having international presence in Singapore and Sri Lanka.

Indian Bank was started in the year 1907, started by V. KrishnaswamyIyer. The founding of the bank was due to the failure of the bank Arbuthnot & Co in 1906 and the Swadeshi Movement. It was the first bank owned and managed by Indians and hence is a symbol of financial freedom.

INTRODUCTION

An activity which may give profits or loss may be called risky due to its unpredictability or uncertainty in future. Risk is present everywhere. There exists many types of financial risk and one of them is credit risk.

Credit risk is defined as the potential that a bank borrower or counterparty will fail to meet its obligations in accordance with agreed terms. In simple words, it means how much customer deviates from his/her promises with the bank.

Ideally when a customer takes loan from the bank, he/she has to repay the loan in periodic installments. But when customer fails to repay any agreed installment, credit risk arises. Because of credit risk, bank suffers huge losses. Because of huge credit risk, bank may get bankrupt. So, its very important to study and minimize this risk.

We can use the previous years data about the customers and their loan transactions to measure this credit risk and can create some algorithm to avoid this credit risk in the future. This data contains the dependent variable which is discrete in nature and hence we need to use standard classification technique like Decision Tree classifier instead of regression analysis.

Decision tree is a graphical representation of possible solutions to a decision based on certain conditions. Decision tree is a type of supervised learning algorithm that is used for classification problems. In this technique, we split the entire population or sample into two or more homogeneous sets based on most significant differentiator in input variables.

SCOPE AND OBJECTIVE

The scope of this project is to construct efficient and accurate models for minimizing the credit risk and maximizing the bank, and to predict whether to give a loan to any customer or not. It also aims in sorting and finding the variables which affects the credit risk of the bank, according to their magnitude. Finally, this project also aims at efficient and easy way to study the prices of the stocks and analyzing the relation between bank's operational variables using Regression analysis.

The aim of this project is to study about the Credit Risk Management, analyzing the big data of the customers and their loan details using standard classification algorithm like decision trees and predicting whether to give a loan to any customer or not. It also includes the analysis of some bank operational variables using Regression analysis and an easy and simple approach to study the prices of the stocks of any company.

CREDIT RISK MANAGEMENT

Credit risk refers to the probability of loss due to lack of a borrower to make payments on any debt. Credit Risk Management is the practice to mitigate these losses by understanding the adequacy of capital reserves and loan losses both a bank in a given time - a process that has long been a challenge for financial institutions.

Credit risk is defined simply as the potential of a bank borrower or counterparty without fulfilling their obligations under the agreed terms. In other words, the credit risk is defined as the risk that no interest or principal or both will be paid as promised and is estimated by observing the proportion of assets that are below standard. Credit risk is assumed by all lenders and will lead to serious problems if it is excessive. For most banks, loans are the largest and most obvious source of credit risk. Creidit risk department of any financial institution needs to be very active throughout the year. They need to carefully analyze every transactions of customers with the institution. If suppose an increasing number of borrowers do not pay back loans as agreed, the bank's profit margin will decrease. This in turn may cause the bank to increase the interest rates and decrease in the amount of loans given to the customers. There are two variants of credit risk are discussed below -

(1) Counterparty risk: This is a variant of credit risk and is related to the failure of trading partners because of the negative and or inability to perform counterpart.

(2) Country Risk: The type of credit risk where failure of a borrower or counterparty arises because of the limitations or restrictions imposed by a country.

Credit risk depends on both external and internal factors. The main **Internal factors** include - The deficiency in credit policy and loan portfolio management. Deficiency in assessing the financial situation of the borrower before lending, Excessive reliance on collateral, Bank failure in the post-sanction monitoring, etc. The main **external factors** include – The state of the economy, Fluctuations in commodity prices, exchange rates and interest rates, etc.

REGRESSION ANALYSIS

A regression analysis is one of the statistical process to estimate relationships between variables. It includes methods for modeling and analyzing several variables, when the focus is on the relationship between one dependent variable and one or more independent variables. More specifically, regression analysis helps to understand how the typical value of the dependent variable get changed when one of the independent variables is varied, while the other independent variables are held fixed.
In general, regression analysis estimates the conditional expectation of the dependent variable given the independent variables - that is, the average value of the dependent variable when the independent variables are fixed. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data.

Linear Regression

In a cause and effect relationship, the independent variable is the cause, and the dependent variable is the effect. Least squares linear regression is a method for predicting the value of a dependent variable Y, based on the value of an independent variable X.

Least Squares Regression Line

Linear regression finds the straight line, called the **least squares regression line** or LSRL, that best represents observations in a bivariate data set. Suppose *Y* is a dependent variable, and *X* is an independent variable. The population regression line is:

$$Y = B_0 + B_1X$$

where B_0 is a constant, B_1 is the regression coefficient, X is the value of the independent variable, and Y is the value of the dependent variable.

Given a random sample of observations, the population regression line is estimated by:

$$\hat{y} = b_0 + b_1x$$

where b_0 is a constant, b_1 is the regression coefficient, x is the value of the independent variable, and ŷ is the *predicted* value of the dependent variable.

Defining a Regression Line

Normally, you will use a computational tool - a software package (for example Excel) - to find b_0 and b_1. You enter the values for X and Y into your program or graphing calculator and the it solves for each parameter.

In the unlikely event that you find yourself on a desert island without a computer or a graphing calculator, you can solve for b_0 and b_1 "by hand". Here are the equations.

$$b_1 = \Sigma [(x_i - x)(y_i - y)] / \Sigma [(x_i - x)^2]$$
$$b_1 = r * (s_y / s_x)$$
$$b_0 = y - b_1 * x$$

where b_0 is the constant in the regression equation, b_1 is the regression coefficient, r is the correlation between x and y, x_i is the *X* value of observation *i*, y_i is the *Y* value of observation *i*, x is the mean of*X*, y is the mean of *Y*, s_x is the standard deviation of *X*, and s_y is the standard deviation of *Y*

Regression Line Properties

When the regression parameters (b_0 and b_1) are defined as described above, the regression line has the following properties.

- The line minimizes the sum of squared differences between observed values (the *y* values) and predicted values (the ŷ values computed from the regression equation).
- The regression line passes through the mean of the *X* values (x) and through the mean of the *Y* values (y).
- The regression constant (b_0) is equal to the y-intercept of the regression line.

- The regression coefficient (b_1) is the average change in the dependent variable (*Y*) for a 1-unit change in the independent variable (*X*). It is the slope of the regression line.

The least squares regression line is the only straight line that has all of these properties.

The **coefficient of determination** (denoted by $\mathbf{R^2}$) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable. It is also denoted by **r**.

- The coefficient of determination ranges from **0 to 1**.
- An $\mathbf{R^2}$ **of 0** means that the dependent variable cannot be predicted from the independent variable while an $\mathbf{R^2}$ **of 1** means the dependent variable can be predicted without error from the independent variable.
- An $\mathbf{R^2}$ **between 0 and 1** indicates the extent to which the dependent variable is predictable. An R^2 of 0.30 means that 30 percent of the variance in *Y* is predictable from *X*; an R^2 of 0.90 means that 90 percent is predictable; and so on.
- The formula to calculate coefficient of determination is given by as follows :

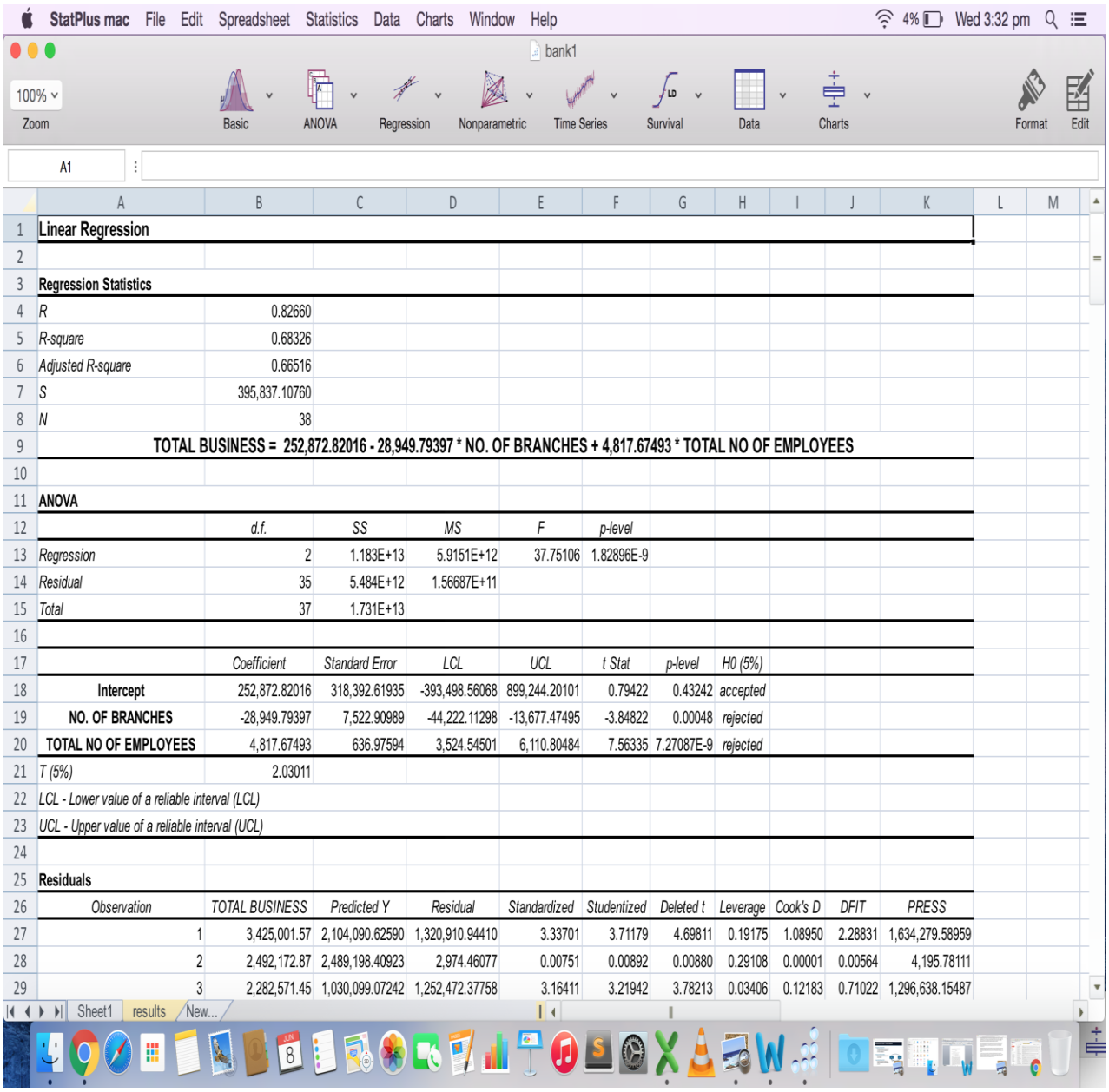
$$\mathbf{r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}}$$

The standard error of the regression line also denoted by SE, is a measure of the average amount that the regression equation over or under predicts. The higher the coefficient of determination, the lower the standard error, and the more accurate predictions are likely to be.

	ZONES	TOTAL BUSINESS	% OF BUSINESS CONTRIBUTED BY THE ZONE	NO. OF BRANCHES	TOTAL NO OF EMPLOYEES
1	MUMBAI	3425001.57	12.38	82	877
2	CHENNAI(SOUTH)	2492172.87	9.01	89	999
3	NEW DELHI	2282571.45	8.25	74	606
4	CHENNAI(NORTH)	1752359.43	6.34	81	891
5	KOLKATA	1151497.91	4.16	85	822
6	BANGALORE	1099620.61	3.98	69	526
7	COIMBATORE	985003.88	3.56	90	708
8	AHMEDABAD	950310.36	3.44	75	550
9	HYDERABAD	943510.54	3.41	58	428
10	MADURAI	641314.97	2.32	94	685
11	KUMBAKONAM	624584.29	2.26	91	667
12	TRICHY	601509.06	2.17	76	595
13	SALEM	598749.63	2.16	78	579
14	PUNE	560775.45	2.03	59	411
15	CUDDALORE	551156.00	1.99	66	513
16	THIRUVANANTHAPU	536805.55	1.94	71	482
17	CHITTOOR	524973.98	1.90	71	505
18	VISAKHAPATNAM	521828.83	1.89	72	544
19	VIJAYAWADA	520448.77	1.88	83	572
20	DHARMAPURI	508845.35	1.84	65	493
21	TIRUNELVELI	477385.68	1.73	72	540

Basis for comparison
Variables to be compared

(The above is the data provided by the Indian Bank regarding the values of total business in different zones of India)



StatPlus mac File Edit Spreadsheet Statistics Data Charts Window Help — 4% Wed 3:32 pm

bank1

Zoom | Basic | ANOVA | Regression | Nonparametric | Time Series | Survival | Data | Charts | Format | Edit

A1

Linear Regression

Regression Statistics

R	0.82660
R-square	0.68326
Adjusted R-square	0.66516
S	395,837.10760
N	38

TOTAL BUSINESS = 252,872.82016 - 28,949.79397 * NO. OF BRANCHES + 4,817.67493 * TOTAL NO OF EMPLOYEES

ANOVA

	d.f.	*SS*	*MS*	*F*	*p-level*
Regression	2	1.183E+13	5.9151E+12	37.75106	1.82896E-9
Residual	35	5.484E+12	1.56687E+11		
Total	37	1.731E+13			

	Coefficient	*Standard Error*	*LCL*	*UCL*	*t Stat*	*p-level*	*H0 (5%)*
Intercept	252,872.82016	318,392.61935	-393,498.56068	899,244.20101	0.79422	0.43242	*accepted*
NO. OF BRANCHES	-28,949.79397	7,522.90989	-44,222.11298	-13,677.47495	-3.84822	0.00048	*rejected*
TOTAL NO OF EMPLOYEES	4,817.67493	636.97594	3,524.54501	6,110.80484	7.56335	7.27087E-9	*rejected*
T (5%)	2.03011						

LCL - Lower value of a reliable interval (LCL)

UCL - Upper value of a reliable interval (UCL)

Residuals

Observation	*TOTAL BUSINESS*	*Predicted Y*	*Residual*	*Standardized*	*Studentized*	*Deleted t*	*Leverage*	*Cook's D*	*DFIT*	*PRESS*
1	3,425,001.57	2,104,090.62590	1,320,910.94410	3.33701	3.71179	4.69811	0.19175	1.08950	2.28831	1,634,279.58959
2	2,492,172.87	2,489,198.40923	2,974.46077	0.00751	0.00892	0.00880	0.29108	0.00001	0.00564	4,195.78111
3	2,282,571.45	1,030,099.07242	1,252,472.37758	3.16411	3.21942	3.78213	0.03406	0.12183	0.71022	1,296,638.15487

Sheet1 | results | New...

The above screenshot is the result of the performed regression analysis.

Findings of the performed regression analysis :

Variable	p-values	Ho (5% significance)
No. of Branches	0.00048	rejected
Total no. of employees	7.27087E-9	rejected

Effect of Total number of branches and Total number of employees on Total business value

Value of correlation is **0.82660.**

Coefficient of Determination which is R square is 0.68 which implies that 68% of the variation in the variable total business is explained by the two variables no. of branches and total number of employees. It also means that total business depends on other factors as well which explains the remaining 32% of the variation.

As the p value of both the independent variable is less than 5% significance level both the coefficients are significant indicating that total business depends on both no. of branches and total number of employees (using t test).

TOTAL BUSINESS is negatively correlated with NO. OF BRANCHES. It also means that if we increase NO. OF BRANCHES by 1, TOTAL BUSINESS will decrease by 28,949.79397. But, it feels little strange because here we did not consider the other factors like in which city of the branch, demographic properties of the branch like location of the branch, population of the society near the branch area.

For example, consider 2 examples from our data,

Mumbai has total business of 3425001.57 units with number of branches being 82 and **Kumbakonam** has total business of 624584.29 units with total number of branches being 91. Mumbai has higher amount of total business as compared to Kumbakonam despite being having lower number of branches. This may be because of number of factors that we ignore like population, living standard, popularity of bank, competitors of the questioned bank and many more.

TOTAL BUSINESS is positively correlated with TOTAL NO OF EMPLOYEES. It also means if we increase TOTAL NO OF EMPLOYEES by 1 unit, TOTAL BUSINESS will increase by 4817.67493 units . But again, it cannot be said like this that if numbers of employees are large then total business will be more. It depends on other factors as well. For example, Mumbai has total business of 3425001.57 units with total number of employees being 877 and Chennai (South) has total business 2492172.87 units with total number of employees being 667.

Prediction of Credit Risk using Loan Data

Credit Risk is the risk that arises when borrower fails to meet its obligations in accordance with the agreed terms. In simple words, it measures how much the borrowers deviates from his/her promises. In Credit Risk Management, Default is the condition when customer fails to repay his/her loan. And Defaulter means the customer who fails to repay the loan. These both terms Default and Defaulter are the most used terms in any credit risk department or division of any bank. It's very necessary to determine the risk which involves reviewing the borrower's past credit history and the income earned.

There are several types of risks can affect a bank. However, counterparty risk or credit risk is both the first, the most dangerous and common risk face a financial institution. In general, the credit risk is defined as the risk that a borrower it defaults: they are not able to keep their promise to pay timely interest payments or repay principal at maturity. Credit risk considers various number of factors to determine the likelihood that a borrower would deviates his/her promises with the loan terms.

Problem Statement :

Based on the previous data of customer and bank details and whether customer paid his/her loan to the bank or not, our bank just want to predict whether to give loan to a new customer or not.

THE DATA

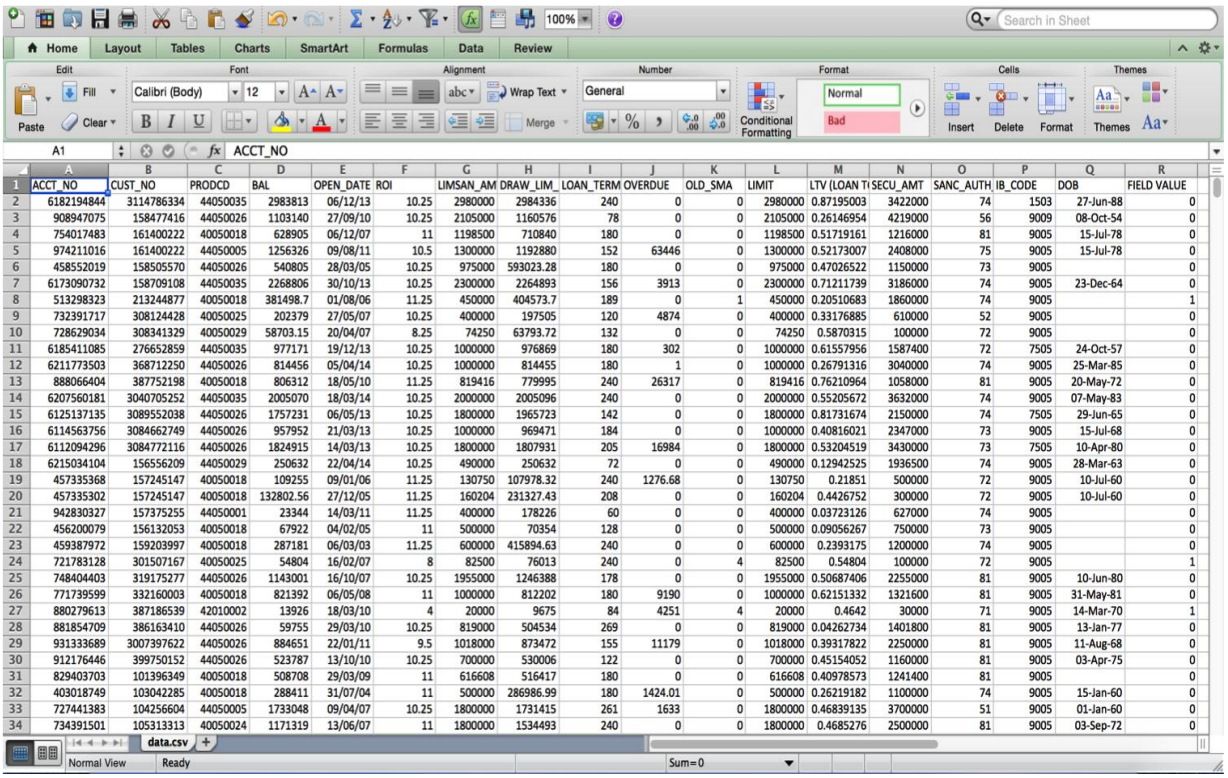
IBGA	ACCT_NO	CUST_NO	PRODCD	BAL	OPEN_DATE	FIELD VALU	ROI	INT_RATE	LIMSAN_AMT	DRAW_LIM_AMT	LOAN_TERM	OVERDUE
V023	6.18E+09	3114786334	4.4E+07	2983813	06-12-2013	0	10.25	10.25	2980000	2984336	240	0
G008	908947075	158477416	4.4E+07	1103140	27-09-2010	0	10.25	10.25	2105000	1160576	78	0
N019	754017483	161400222	4E+07	628905	06-12-2007	0	11	11	1198500	710840	180	0
N019	974211016	161400222	4.4E+07	1256326	09-08-2011	0	10.5	10.5	1300000	1192880	152	63446
U016	458552019	158505570	4.4E+07	540805	28-03-2005	0	10.25	10.25	975000	593023.28	180	0
A074	6.17E+09	158709108	4.4E+07	2268806	30-10-2013	0	10.25	10.25	2300000	2264893	156	3913
P042	513298323	213244877	4E+07	381499	01-08-2006	1	11.25	11.25	450000	404573.7	189	0
T091	732391717	308124428	4E+07	202379	27-05-2007	0	10.25	10.25	400000	197505	120	4874
W004	728629034	308341329	4E+07	58703.2	20-04-2007	0	8.25	8.25	74250	63793.72	132	0
S062	6.18E+09	276652859	4.4E+07	977171	19-12-2013	0	10.25	10.25	1000000	976869	180	302
S012	6.21E+09	368712250	4.4E+07	814456	05-04-2014	0	10.25	10.25	1000000	814455	180	1
A074	888066404	387752198	4E+07	806312	18-05-2010	0	11.25	11.25	819416	779995	240	26317
B086	6.20E+09	3040705252	4.4E+07	2005070	18-03-2014	0	10.25	10.25	2000000	2005096	240	0
P132	6.12E+09	3089552038	4.4E+07	1757231	06-05-2013	0	10.25	10.25	1800000	1965723	142	0
K165	6.11E+09	3084662749	4.4E+07	957952	21-03-2013	0	10.25	10.25	1000000	969471	184	0
M035	6.11E+09	3084772116	4.4E+07	1824915	14-03-2013	0	10.25	10.25	1800000	1807931	205	16984
M033	6.21E+09	156556209	4.4E+07	250632	22-04-2014	0	10.25	10.25	490000	250632	72	0
K144	457335368	157245147	4E+07	109255	09-01-2006	0	11.25	11.25	130750	107978.32	240	1276.68
K144	457335302	157245147	4E+07	132803	27-12-2005	0	11.25	11.25	160204	231327.43	208	0
M154	942830327	157375255	4.4E+07	23344	14-03-2011	0	11.25	11.25	400000	178226	60	0
K132	456200079	156132053	4E+07	67922	04-02-2005	0	11	11	500000	70354	128	0
T068	459387972	159203997	4E+07	287181	06-03-2003	0	11.25	11.25	600000	415894.63	240	0
F004	721783128	301507167	4E+07	54804	16-02-2007	1	8	8	82500	76013	240	0
K160	748404403	319175277	4.4E+07	1143001	16-10-2007	0	10.25	10.25	1955000	1246388	178	0
P083	771739599	332160003	4E+07	821392	06-05-2008	0	11	11	1000000	812202	180	9190
S177	880279613	387186539	4.2E+07	13926	18-03-2010	1	4	4	20000	9675	84	4251
N110	881854709	386163410	4.4E+07	59755	29-03-2010	0	10.25	10.25	819000	504534	269	0
G060	931333689	3007397622	4.4E+07	884651	22-01-2011	0	9.5	9.5	1018000	873472	155	11179
T017	912176446	399750152	4.4E+07	523787	13-10-2010	0	10.25	10.25	700000	530006	122	0
A132	829403703	101396349	4E+07	508708	29-03-2009	0	11	11	616608	516417	180	0

	A	B
1	CUST_ID	NPA_DATE
2	163207590	21-SEP-13
3	213244877	31-MAR-14
4	367406720	12-JAN-12
5	3005741986	30-JUN-13
6	329183707	27-AUG-13
7	348198325	15-MAY-13
8	3038284873	09-OCT-13
9	110122322	31-MAR-95
10	3016711029	18-OCT-13
11	263350998	25-APR-13
12	309142866	27-DEC-13
13	286867432	01-MAR-96
14	395654645	31-MAR-14
15	326285248	30-SEP-13
16	3070435824	19-SEP-13
17	395078567	30-JUN-13
18	237014572	13-FEB-14
19	135751216	31-MAR-14
20	357466577	12-NOV-13
21	258874259	30-SEP-13
22	3046848344	30-SEP-13
23	3014494209	30-JUN-13
24	3048185789	27-NOV-13
25	353558699	10-JUL-13

This data contains detail about the customers, their previous loan transactions and some details about their bank. This data constitutes of 17 independent variables like Account number, customer id, rate of interest value, old sma, loan to value ratios, date of birth of the customers, opening date of the consumer's transaction, account balance, product code, security amounts, ib code, ibga code, loan term, limsan amounts etc.

There is one file which contains detail of NPA, whether customer defaulted their loan or not. Then comparing both these files, we need to check whether

customers of the first file are present in the second file. Based on that, if customer of the first file is present in second file, then assign field value of 1 and if customer of the first file is not present in the second file, then assign field value of 0. After creating column of field values in the first file, our first file is ready. Then we need to clean the data, to make it ready for the analysis. This is the most time consuming process in any data analysis. There were many missing values in the date of birth section and opening date section. So, we need to remove those values or else we need to fill those missing values with some techniques. In this way, we can make the data final for the analysis. In this data, some missing values were filled while some where removed at all. Also, some of the columns which were not useful where removed and a new file of data which need to be analyzed is made. Here is the image of the final prepared data.



ACCT_NO	CUST_NO	PRODCD	BAL	OPEN_DATE	ROI	LIMSAN_AM	DRAW_LIM_	LOAN_TERM	OVERDUE	OLD_SMA	LIMIT	LTV (LOAN TI	SECU_AMT	SANC_AUTH	IB_CODE	DOB	FIELD VALUE
6182194844	3114786334	44050035	2983813	06/12/13	10.25	2980000	2984336	240	0	0	2980000	0.87195003	3422000	74	1503	27-Jun-88	0
908947075	158477416	44050026	1103140	27/09/10	10.25	2105000	1160576	78	0	0	2105000	0.26146954	4219000	56	9009	08-Oct-54	0
754017483	161400222	40050018	628905	06/12/07	11	1198500	710840	180	0	0	1198500	0.51719161	1216000	81	9005	15-Jul-78	0
974211016	161400222	44050005	1256326	09/08/11	10.5	1300000	1192880	152	63446	0	1300000	0.52173007	2408000	75	9005	15-Jul-78	0
458552019	158505570	44050026	540805	28/03/05	10.25	975000	593023.28	180	0	0	975000	0.47026522	1150000	73	9005		0
6173090732	158709108	44050035	2268806	30/10/13	10.25	2300000	2264893	156	3913	0	2300000	0.71211739	3186000	74	9005	23-Dec-64	0
513298323	213244877	40050018	381498.7	01/08/06	11.25	450000	404573.7	189	0	1	450000	0.20510683	1860000	74	9005		1
732391717	308124428	40050025	202379	27/05/07	10.25	400000	197505	120	4874	0	400000	0.33176885	610000	52	9005		0
728629034	308341329	40050029	58703.15	20/04/07	8.25	74250	63793.72	132	0	0	74250	0.5870315	100000	72	9005		0
6185411085	276652859	44050035	977171	19/12/13	10.25	1000000	976869	180	302	0	1000000	0.61557956	1587400	72	7505	24-Oct-57	0
6211773503	368712250	44050026	814456	05/04/14	10.25	1000000	814455	180	1	0	1000000	0.26791316	3040000	74	9005	25-Mar-85	0
888066404	387752198	40050018	806312	18/05/10	11.25	819416	779995	240	26317	0	819416	0.76210964	1058000	81	9005	20-May-72	0
6207560181	3040705252	44050035	2005070	18/03/14	10.25	2000000	2005096	240	0	0	2000000	0.55205672	3632000	74	9005	07-May-83	0
6125137135	3089552038	44050026	1757231	06/05/13	10.25	1800000	1965723	142	0	0	1800000	0.81731674	2150000	74	7505	29-Jun-65	0
6114563756	3084662749	44050026	957952	21/03/13	10.25	1000000	969471	184	0	0	1000000	0.40816021	2347000	73	9005	15-Jul-68	0
6112094296	3084772116	44050026	1824915	14/03/13	10.25	1800000	1807931	205	16984	0	1800000	0.53204519	3430000	73	7505	10-Apr-80	0
6215034104	156556209	44050029	250632	22/04/14	10.25	490000	250632	72	0	0	490000	0.12942525	1936500	74	9005	28-Mar-63	0
457335368	157245147	40050018	109255	09/01/06	11.25	130750	107978.32	240	1276.68	0	130750	0.21851	500000	72	9005	10-Jul-60	0
457335302	157245147	40050018	132802.56	27/12/05	11.25	160204	231327.43	208	0	0	160204	0.4426752	300000	72	9005	10-Jul-60	0
942830327	157375255	44050001	23344	14/03/11	11.25	400000	178226	60	0	0	400000	0.03723126	627000	74	9005		0
456200079	156132053	40050018	67922	04/02/05	11	500000	70354	128	0	0	500000	0.09056267	750000	73	9005		0
459387972	159203997	40050018	287181	06/03/03	11.25	600000	415894.63	240	0	0	600000	0.2393175	1200000	74	9005		0
721783128	301507167	44050025	54804	16/02/07	8	82500	76013	240	0	4	82500	0.54804	100000	72	9005		1
748404403	319175277	44050026	1143001	16/10/07	10.25	1955000	1246388	178	0	0	1955000	0.50687406	2255000	81	9005	10-Jun-80	0
771739599	332160003	40050018	821392	06/05/08	11	1000000	812202	180	9190	0	1000000	0.62151332	1321600	81	9005	31-May-81	0
880279613	387186539	42010002	13926	18/03/10	4	20000	9675	84	4251	4	20000	0.4642	30000	71	9005	14-Mar-70	1
881854709	386163410	44050026	59755	29/03/10	10.25	819000	504534	269	0	0	819000	0.04262734	1401800	81	9005	13-Jan-77	0
931333689	3007397622	44050026	884651	22/01/11	9.5	1018000	873472	155	11179	0	1018000	0.39317822	2250000	81	9005	11-Aug-68	0
912176446	399750152	44050026	523787	13/10/10	10.25	700000	530006	122	0	0	700000	0.45154052	1160000	81	9005	03-Apr-75	0
829403703	101396349	40050018	508708	29/03/09	11	616608	516417	180	0	0	616608	0.40978573	1241400	81	9005		0
403018749	103042285	40050018	288411	31/07/04	11	500000	286986.99	180	1424.01	0	500000	0.26219182	1100000	74	9005	15-Jan-60	0
727441383	104256604	44050005	1733048	09/04/07	10.25	1800000	1731415	261	1633	0	1800000	0.46839135	3700000	51	9005	01-Jan-60	0
734391501	105313313	40050024	1171319	13/06/07	11	1800000	1534493	240	0	0	1800000	0.4685276	2500000	81	9005	03-Sep-72	0

Approach required to analyze the data :

There are mainly two types of problems in data analysis. First one is Regression and other one is classification. The type of approach needed to analyze the data depends on the nature of the output. Using Regression, we predict output which takes only continuous values. For example, predicting price of the houses using some number of variables. In Classification, we predict the output which takes only discrete values. For example, in our case it is whether to give a loan to customers or not.

Lets see an example here. Price of a house depending on the size and say suppose location of the house, can be some numerical value which is continuous in nature, this relates to regression. In, we are trying to predict the results within a continuous output, which means we are trying to assign input variables for any continuous function. In the regression in a classification problem, rather we are trying to predict the results in a discrete output.

The main difference between the tree classification and regression tree is the dependent variable. For the classification tree, the dependent variables are categorical, while the regression tree is dependent numerical variables. The classification of trees also have a fixed amount of unordered values, whereas regression tree have any of the values still ordered discrete values or indiscreet. A regression tree is constructed in order to adjust a system of determining regression for each branch in a way that the expected output value appears. Furthermore, classification tree branches as a dependent variable determined by the previous node derivative.

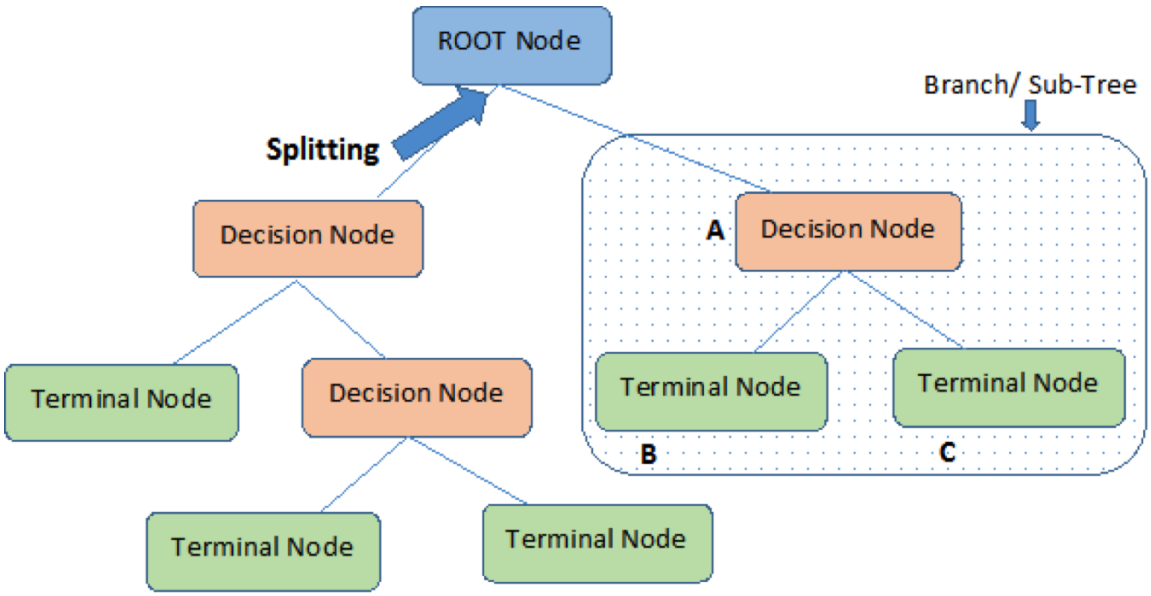
So here in this case our output is the field value column that is 0 or 1. So, we see output is in the discrete form. So, we see it is the problem of classification. We need to apply any standard classification technique like decision tree. From the original data, we removed total 249 rows of data as they were containing missing values which needs to be removed to make our data eligible for analysis.

DECISION TREES

A decision tree is a graphical representation of the possible solutions to a decision based on certain conditions. A decision tree is called, as it begins with a single box (or root), then branches into a number of solutions, like a tree. It is one of the machine learning algorithms used for classification. It helps us in making the best decisions on the basis of existing information and best guesses. It allows you to fully analyze the possible sequences of a decision.

Decision trees are useful, not only because they are graphics that help you see what you are thinking, but also because making a decision tree requires a process of systematic and documented thought. Often, the greatest limitation of our decision is that we can only select from the known alternatives. Decision trees help to formalize the process of exchange of ideas so that we can identify potential solutions. A decision tree is a graph uses a branching method to illustrate all possible outcomes of a decision.

Decision trees can be drawn by hand or created with a graphics program or specialized software. Informally, decision trees are useful to focus the discussion when a group must make a decision. The variables in a decision tree usually represented by rectangles and circles. Unlike the other models of decision making, the decision tree contains all the possible alternatives and trace each alternative to its conclusion at a single glance, allowing easy comparison between various alternatives. The use of separate nodes to denote decisions by the user, uncertainties defined, and at the end of the process clarity and transparency lends itself to trees process. Decision trees decomposes the data in an easy to understand illustrations, based on easily understandable rules for loved human and SQL programs. Decision trees also allow data classification without calculation, can handle both continuous and categorical variables, and provide a clear indication of the most important fields used for the prediction and classification. It is the best predictive model. They also provides a graphical illustration of the problem and alternatives in a simple and easy to understand format that does not require explanation.



Note:- A is parent node of B and C.

Terminologies used in Decision Trees :

- ***ROOT Node***: It is the most basic node of any decision tree. It is also the starting node of any decision tree. It represents the entire sample or population and this further gets divided into two or more homogeneous sets of data.
- ***SPLITTING***: It is a process of dividing a node into two or more sub-nodes. It occurs when any node gets distributed into further nodes. At every decision node, we can say that splitting occurs.
- ***Decision Node***: When a node splits into further sub-nodes, then it is called decision node. It is called as decision node because at that node, some decision has to be taken place. At every decision node, phenomenon of splitting occurs.
- ***Leaf/ Terminal Node***: Nodes do not split is called Leaf node. They are also called as terminal node. These are the nodes where no further splitting occurs. In decision tree classifier, these leaf nodes gives the final decision to be taken based on the decision tree algorithm.
- ***Parent and Child Node***: A node, which is further divided into sub-nodes is called parent node of those sub-nodes where as these sub-nodes are the child of those parent node. Leaf nodes don't have any child node, as they are not splitted any further.

CHAID ANALYSIS

Chi-square Automatic Interaction Detector (CHAID) was a technique created by Gordon V. Kass in 1980. Chaid it is a tool used to discover the relationship between the variables. This chaid analysis builds a predictive model or a tree, to help determine how variables combine to better explain the results given in the dependent variable. In simple words, it is a form of **analysis** that determines how variable best combine to explain the outcome in a given dependent variable. The model can be used in cases of market penetration, predicting and interpreting responses or a multitude of other research problems.

Chaid analysis is especially useful for data expressing categorized values instead of continuous values. For this kind of data some common statistical tools such as regression are not applicable and Chaid analysis is a perfect tool to discover the relationship between variables.

One of the outstanding advantages of CHAID analysis is that it can visualize the relationship between the target (also called as dependent) variable and the related factors with a tree image like we used here. This makes it very easy to interpret the results and helps easily in taking decisions. Chaid analysis is used to build a predictive model to outline a specific customer group or segment (group) - for example, loan worthy cutomers or customers who deserves the loan.

Chaid uses predictor variables like those 17 variables in our excel data file to divide the sample into a number of subgroups that share similar characteristics called "decision tree". Like here in our case, two major segments will be like one those who deserves the loan and other one who does not deserves the loan. These subgroups allow prediction of group membership - what are the characteristics of satisfied customers - as well as predicting the value of group membership in each division - how satisfied customers are in each branch of the "decision tree ".

Code of the Project

sir1.py

```
21
22  days = []
23  for d in od:
24      if type(d) == float:
25          days.append(d)
26      else:
27          try:
28              dt = datetime.datetime.strptime(d, "%d/%m/%y")
29              if dt.year > 2014:
30                  dt = dt - datetime.timedelta(weeks=52*100)
31                  days_in_days = (ref - dt).days
32                  days.append(days_in_days)
33              else:
34                  days_in_days = (ref - dt).days
35                  days.append(days_in_days)
36          except:
37              days.append(None)
38
39  data['AGE'] = pandas.Series(age)
40  data['DAYS'] = pandas.Series(days)
41
42  data = data.drop('AGE', 1)
43  data = data.drop('DOB', 1)
44
45  new_data = data.dropna()
46
47  train_X = new_data[["BAL", "ROI", "LIMSAN_AMT", "DRAW_LIM_AMT", "LOAN_TERM", "OVERDUE", "OLD_SMA", "SECU_AMT", "DAYS"]].values
48  train_Y = new_data["FIELD VALUE"].values
49
50  import sklearn
51  from sklearn.tree import DecisionTreeClassifier
52  clf = DecisionTreeClassifier()
53  clf = clf.fit(train_X, train_Y)
54  from sklearn.externals.six import StringIO
55  import pydot
56  from sklearn import tree
57  dot_data = StringIO()
58
59  clf = DecisionTreeClassifier(max_depth=4)
60  clf = clf.fit(train_X, train_Y)
61  dot_data = StringIO()
62  tree.export_graphviz(clf, out_file=dot_data, feature_names=["BAL", "ROI", "LIMSAN_AMT", "DRAW_LIM_AMT", "LOAN_TERM", "OVERDUE", "OLD_SMA", "SECU_AMT", "DAY
63  graph = pydot.graph_from_dot_data(dot_data.getvalue())
64  graph[0].write_pdf("dtree4.pdf")
65
66
67
68
69
```

Line 1, Column 1 | Spaces: 4 | Python

Meaning of the above code

First of all, to use the decision tree algorithm and to plot these decision trees, install lot of machine learning packages and graphic packages. There is need to install numpy, scipy, pandas, scikit-learn, pydot and some more graphic packages. These all are the packages which are used to plot the decision trees for the data.

Then input the data into some variable which will be used later. Then clean this data by filling missing values with their means and removing some unnecessary rows and colums. Also, some of the missing values can be removed.

Then define one classifier object like **clf** variable defined in the code which will be the classifier object. Then define the training set X and Y. The training set X includes the independent variables (predictors) which are used in the model and the training set Y includes the dependent variable which in this case is the field value variable. Then apply the required algorithm on the classifier object. Here, in this case the algorithm used for the purpose of classification was decision tree. So, apply decision tree classification algorithm on the classifier object which here is clf.

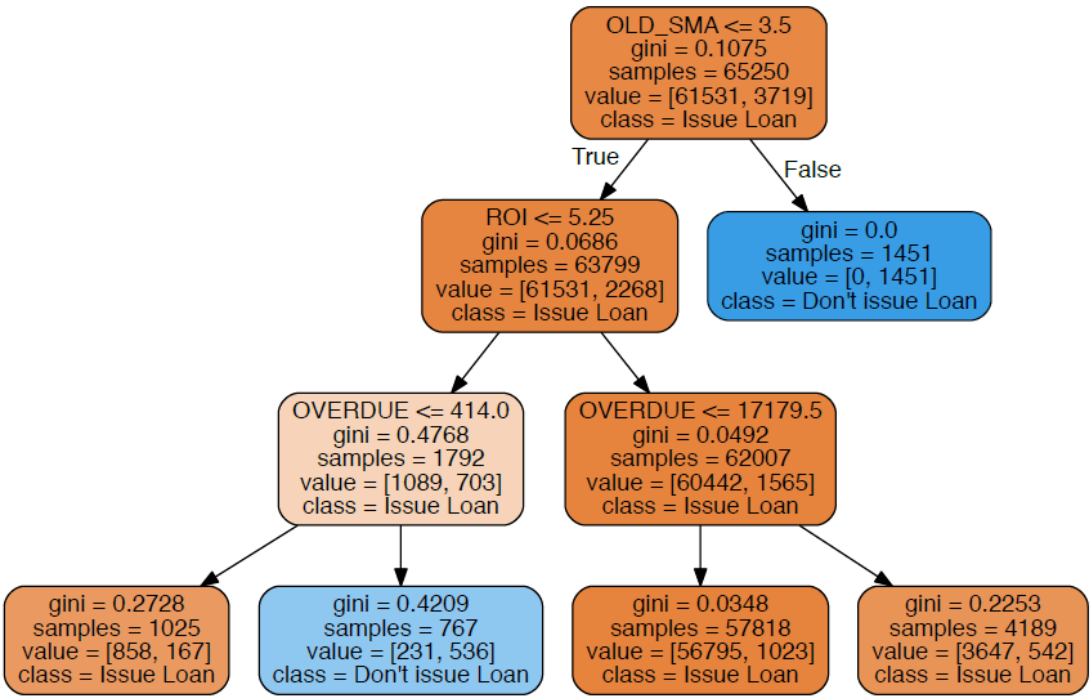
Specify the depth of the decision tree you need. For example, if you need decision tree of depth=3, then pass the parameter max_depth=3 in the decision tree classifier method like shown below.

clf = DecisionTreeClassifier(max_depth=3)

After completion of code, save the code file. After saving, run that code file and follow the output. In this way, whole code for plotting the decision trees for any classification problem can be written.

Output of the code :

Decision tree of depth = 3



This decision tree diagram shows that if value of old_sma is greter than 3.5, then we should not allow customer to take the loan. If old_sma value is less than or equal to 3.5, then we need to see the rate of interest value. If rate of interest is less than or equal to 5.25, then we can allow customer to take the loan form the bank. If rate of interest is greater than 5.25, then we need to check the value of Overdue. In this way, we can take a decision whether to give a loan to a customer or not. Also, in this diagram orange colored shaded box tells us to issue the loan and blue colored shaded box tells us that customer is not eligible for a loan. Darker the shade of any node, more authentic decision we can conclude. Also, each node contains value of gini coefficient. This variable tells us about the strength of the classification that node is trying to tell. Lower the value of gini-coefficient that is as it approaches to 0, stonger and more accurate classification is done. Higher the value of gini-coefficient that is as it approaches to 1, weaker and less

accurate classification is done. The value variable in each node represents an array of 2 elements, in which first element represent the number of people who are eligible for the loan and the second element represents the number of people who are not eligible for the loan. Based on this array, we can predict the accuracy and strength of the classification done by that node.

Confusion Matrix :

A confusion matrix is a table that is often used to describe the performance of a classification model in a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but related terminology can be confusing.
Each row of the matrix represents the instances in an actual class while each column of the matrix represents the instances in a predicted class. Confusion Matrix is mainly used to measure the accuracy and performance of the classification model.

There are mainly two useful terms that can be calculated using any confusion matrix like accuracy, misclassification rate. Let's see each term one by one:

Accuracy : Overall, how often is the classifier correct?

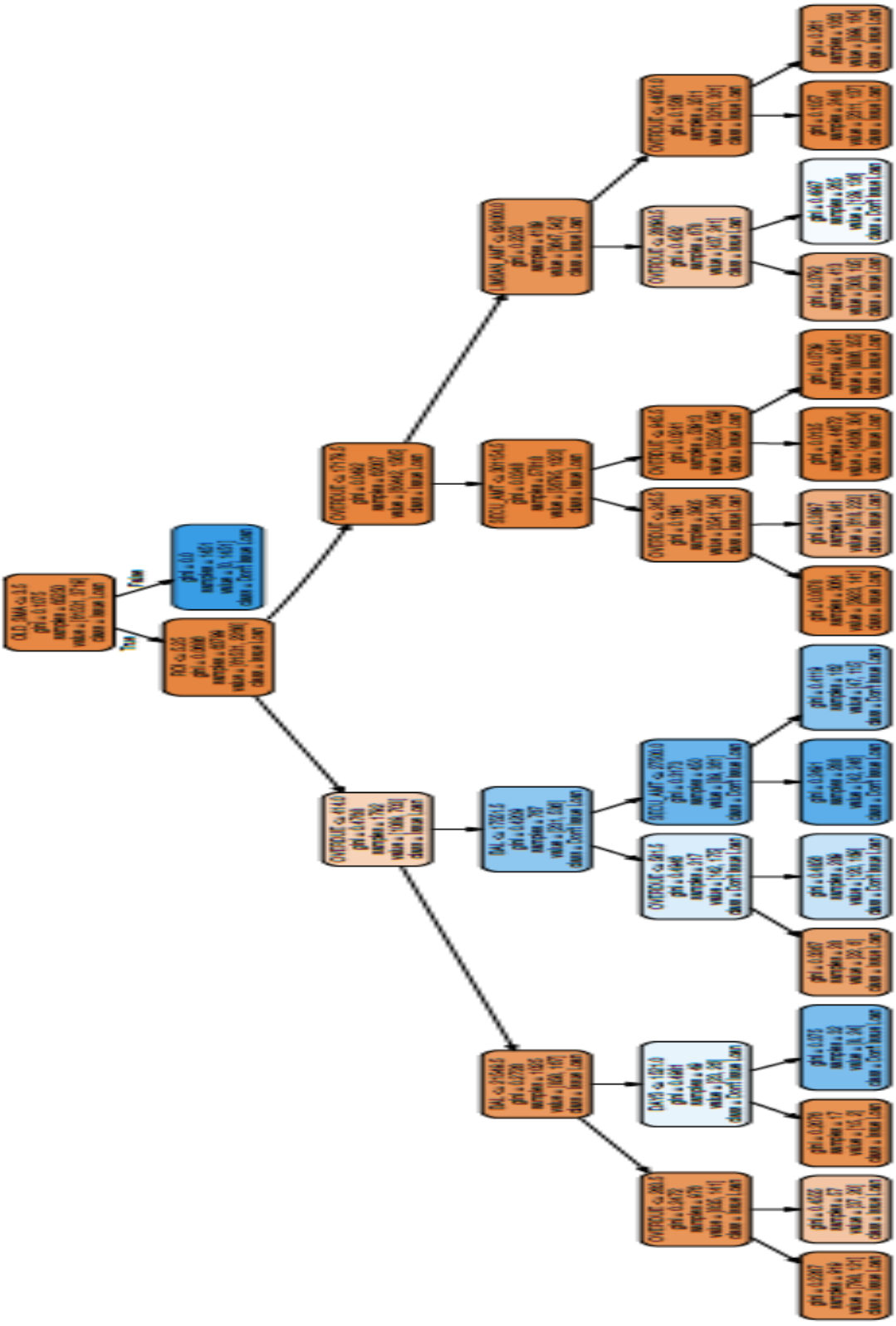
Misclassification Rate: Overall, how often is it wrong?

Confusion Matrix for decision tree whose depth = 3

N = 65250 samples	Predicted: Issue loan	Predicted: Don't issue loan
Actual: Issue loan	61300	231
Actual: Don't issue loan	1732	1987

Accuracy : (61300+1987)/65250 = 0.9699 * 100 = 96.99%. So, we see our classification model is predicting 96.99% of the correct results.
Misclassification Rate: (1732+231)/65250 = 0.03 * 100 = 3%. So, there is 3% misclassification done by this decision tree model.

Decision tree of depth = 5



Confusion Matrix for decision tree whose depth = 5

N = 65250 samples	Predicted: Issue loan	Predicted: Don't issue loan
Actual: Issue loan	61185	346
Actual: Don't issue loan	1578	2141

Accuracy : (61185+2141)/65250 = 97.05%. Hence this decision tree is predicting 97.05% of correct results.
Misclassification rate: (1578+346)/65250 = 2.95%. Hence there is 2.95% misclassification done by this decision tree model.

For a bank, the cost of issuing a loan to a customer who actually doesn't deserve a loan is more than the cost of not issuing a loan to a customer who deserves a loan.

Loss Equation for the above model :

Loss = 1578 * average loss bank suffers from 1 defaulter - 346 * average loss bank suffers from not giving loan to 1 customer who deserves loan

We need to minimize the above loss. This way loss can be calculated and credit risk can be minimized and avoided. We also see that accuracy of the decision tree of depth = 3 model and decision tree of depth = 5 model are almost same about the 97%. But the loss for both the models is different.
Loss is less in the later model (decision tree with depth = 5) when compared to the first model (decision tree with depth = 3). It is because the number of customers which makes bank suffer more loss are greater in the later model when compared to the first model. These numbers are 1578 in the second model and 1732 n the first model. So, decision tree model with depth=5 will be preferred more over the decision tree model with depth=3. Also, the number of predictors are more in decision tree with depth=5 model when compared to the first model. This generates greater flexibility in our decisions.

CONCLUSIONS

In the performed regression analysis of analyzing the effect of total number of branches and Total number of employees on total business value, the result was that Coefficient of Determination which is R square is **0.68** which implies that **68% of the variation** in the variable total business is explained by the two variables no. of branches and total number of employees. This also means 32% of the remaining variation in the variable total business is explained by some other factors.

Using the decision tree obtained as the results, we can predict the final decision whether to give a loan to a customer or not. The main predictors (independent variables) which affects the final decision are Old_Sma, ROI (rate of interest), Overdue, Bal (balance) , Secu_Amt (security amounts) and Days (number of days between opening date and 30-Apr-14. From all of these predictors, we see old_sma comes out to be the most important predictor in taking decision as it comes at top of the deccision tree. On the other hand, value of rate of interest (ROI) and Overdue also matters a lot.

REFERENCES

1) www.indian-bank.com

2) www.google.co.in

3) http://scikit-learn.org/stable/modules/tree.html

4) https://www.youtube.com/watch?v=7gAZoK6kGhM

5) Credit Risk Management - Basic Concepts by Tony Gestel and Baesens