Credit Scoring Process using Banking Detailed Data Store

Meera Rajan
Consultant -Risk Management Mumbai
Info Drive Analytics Pvt. Ltd.

Tulasi. B
Department of Computer Science
Christ University

ABSTRACT

Credit scoring process has become the current popular need of the sectors like Banking, Telecom, and Insurance. The current paper discusses credit scoring for banking Sector. It discusses about Credit Scoring for BASEL II, also to build an integrated solution for it. The framework of credit scoring solution is to enable a bank to build Analytic models for application score or Probability of Default (PD),Loss Given default(LGD), Credit Conversion Factor (CCF). The credit scoring process is integrated with the Credit Risk Management. In this paper the SAS tool named SAS E-Miner is used to perform Credit Scoring using DDS (Detailed Data Store) and SEMMA methodology is applied.

Keywords

Credit Scoring, Logistic Regression, SEMMA, Detailed Data Store, SAS E-miner

1. INTRODUCTION

Credit scoring can be explained as a building a statistical Model to monitor the effect on impact score-based decisions that have an impact on business key variables or key indicators. The output/s obtained from credit scoring techniques aids lenders or banks in the decision making on whether to grant credit to customers or not. DDS (Detailed data Store) is a data mart that provides a single version of truth for Banking Intelligence Solutions. It helps in storing historical data for Model Building and validation purpose and also helps in extracting the same at a particular point in time.

The process meets the requirements

Audits and Data validation

Integration with credit Risk Management

Support integration with third party loan organization.

Audits and Data validation include:

a. Data Integration Documentation that describes the Integration process for Data (ETL process) [2]

b. Rating model development documentation, explains the creation of rating model.

Credit Risk Management is managing the credit risk. Credit risk can be defined as the potential for loss due to failure of a borrower to meet the requirement to repay a loan or debt in accordance to predefined or agreed terms. The credit risk events include Bankruptcy and failure to pay loan installments. Credit Risk Management is involved in seamlessly integrating credit-scoring processes with overall credit portfolio risk assessment.

The credit scoring is one of widely used applications of Data Mining, as it can predict consumer behavior. SEMMA methodology is used for model development. In this paper Logistic Regression model and PCA (Principal Component Analysis) model is generated using SAS E-Miner tool for Credit Scoring.

This paper discussion is about Banking DDS and the scope of the paper is focused on the measures that may be implemented for better tuning of the ETL by reducing the need of rewriting the logic for a job [1]

1.1 Credit Scoring and Data Mining

Credit Scoring is one of the earliest uses of data to predict the behavior of a consumer. It is also considered as one of oldest applications of data mining. Data Mining is defined process of Selecting, Exploring, and Modeling enormously large amount of data for unknown patterns, certain relationships that can be exploited for automated decision making.

1.2 Data Mining and SEMMA Methodology

SEMMA is one of the most widely accepted methodologies of Data Mining. SEMMA is a Data Mining Model used by SAS E-Miner. SEMMA stands for

Table 1: SEMMA Model

Character	**Description**
S - SAMPLE	Sample Data by creating one or more data tables.
E - EXPLORE	Exploring the Data information about the dataset. Search for relationships, anticipated for trends in order to gain understanding of data
M - MODIFY	Modify the data by creating, selecting and transforming the variables to focus on the model selection process.
M - Model	Model the data by using analytical tools .To search for combination that is reliable to acquire desired outcome of the model.
A - Assess	Assess for completing a model, to tune into champion model by back testing etc. Model is compared for certain criteria and once model performs well then it is promoted to the champion model from the challenger model.

2. OBJECTIVES OF CREDIT SCORING

The scope of credit scoring solution in Banks enables them to:

Build analytical models for Application Score

The analytical models are built to generate scorecard used for credit scoring. The analytical models help the bank to get an idea of the behavior of a customer, to understand the relationship and importance of the various parameters or key variables involved in solving a specific problem specification. After building the model it aids in understanding the actual scenario the bank may foresee by giving a loan to a customer.

To produce scores for customers monthly / Quarterly / Annually

The periodic scoring of the customers has become the basic priority of a bank for consistent monitoring of their behavior as potential customers [9]. There are high chances of turning current potential customers of a year to non-potential by the next consecutive year. To retain customer potentiality, a continuous periodic scoring process is required monthly, quarterly and annually.

To monitor the performance of analytical models for regular and Business use

The models built once cannot be presumed to be effective forever. The model behavior certainly changes when there are changes in the input parameters and requirement specifications so on and so forth. Therefore the model has to be monitored regularly for its efficiency.

To increase Customer Acceptance Rate

Many credit-scoring models have low acceptance ratio by customers as they are not very user-friendly, therefore a need to increase customer acceptance rate and achieve high customer appreciation has become mandatory.

To support Retail, Corporate and Collection Scorecards

The credit scoring for Banking must support Retail, Corporate and Collection Scorecard models. (The collection scorecard must try to predict a customer's response to various strategies and policies used for collecting their money).

3. CURRENT ISSUES IN CREDIT SCORING

Credit scoring observes the following current issues:

The source systems are scattered among various different sources like Core Banking, Insurance

Programmers and analysts faced major challenges in collecting and integrating the source data. This takes a lot of time and also never guarantees a single version of the truth. Programmers have to put in a lot of time in collecting, organizing and ensuring accurate data.

The data is inconsistent and also scattered across many solutions

The data is highly inconsistent. For example one the source system may use “M” for male and “F” for female while the other may use “1” and “2”. The total information about a particular customer was scattered, so it took up a lot of time and effort to consolidate and understand data. To know whether a customer has repaid the loan, various solutions like loan payment, treasury, accounts etc. have to be verified and consolidated.

To identify the best model built

The models built on a particular data need to be evaluated. More than one model can be built on the same data, whereas the challenge was to identify the best model amongst all for the same data. For example, the same data can be used to construct models using Decision Trees, Neural network, etc. To identify the best was difficult.

4. PROPOSED SOLUTION FOR CREDIT SCORING

To create a detailed data mart for data population:

A detailed data mart is created that also acts as a standardized data mart for all the banking solutions. The detailed standardized data mart aids in reducing ambiguity for future operations or further processing.

Extendable data mart for further variable additions for predictive modeling:

The data marts available were static in nature, it was not possible to expand it. .To extend either a new table or a new column in future was not possible. The data mart had to be designed from scratch. There is an extensive requirement for Extendable data mart.

To build a Credit Score model from Detailed Data store:

Detailed Data Store reflects information of a single version of truth [3]. The Credit Scoring models built from various sources of data created ambiguity in analyzing the exact facts regarding the customers. To resolve ambiguity and to reflect clarity about the actual fact can be achieved by developing a credit scoring model from detail data store.

5. CREDIT SCORING AND DETAILED DATA STORE

Credit Scoring can be obtained even without using DDS. The major benefit observed using Detailed Data Store is it facilitates integrated approach to build a Model**.**

Integrated Approach of DDS:

DDS follows an integrated approach, as it reflects on a single version of truth. The data from all the external sources is extracted and loaded into the DDS in a standardized procedure and also standardized extraction process. [4].

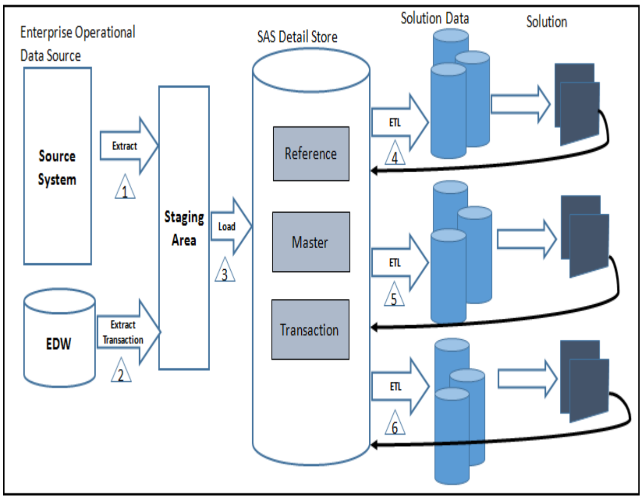


Fig 1: Optimize ETL for the banking DDS

Building And Implementing Integrated Data Models:

The process of building and implementing an integrated Data Model includes the initial step of collecting data from various Enterprise Source Systems, followed by ETL (Extract, Transform, Load) process and after ETL is stored in Banking Detail Data Mart. From Detail Data Mart the respective input, process and output operations depending upon the requirement specifications are executed to achieve desired outputs/reports.

The DDS has other key features apart from integrated approach as stated below:

DDS Modifications: The DDS is an extendable data store that provides dynamic facility of adding fields to the tables. It also gives flexibility to add new data tables.[2]

DDS Exclusive properties:
The major popular use of DDS is that it is very easy to customize. The other data stores available in the industry do not provide an easy way to customize or there is no option to customize at all. This characteristic feature of DDS is a prominent reason for its high acceptability in the industry.

The other exclusive feature is the facility to store historical data for Model building and extracting it whenever needed.

DDS in Credit Scoring:
The data is taken from source data systems, integrated and sent to DDS. The data is then sent from DDS through a foundation layer and stored in ABTs. The Modeling ABT is used to store input data; the scoring ABT is used to store selected data that need to be processed. The data flows from ABTs to respective SAS tools[6] (SAS, SAS E-Miner etc.) and then corresponding reports are generated

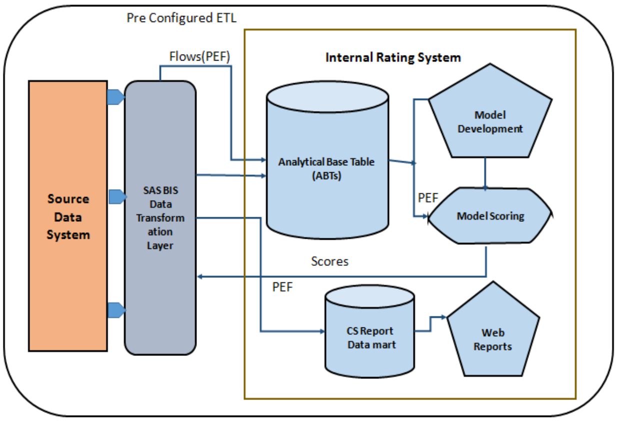


Fig 2: DDS in Credit Scoring-Logical Flow

Analytical Base Tables (ABTs):
The Analytical Base Tables are used for jobs specified below:

It consists of a large list of predefined variables and outcomes.

It has the ability to select a subset of variables and one outcome.

It has the ability to create new variables.

There are two types of ABTs:
Modeling ABT:
It is used for building Analytical Models in SAS E-Miner. It is more for an input operation to store required input data for processing.

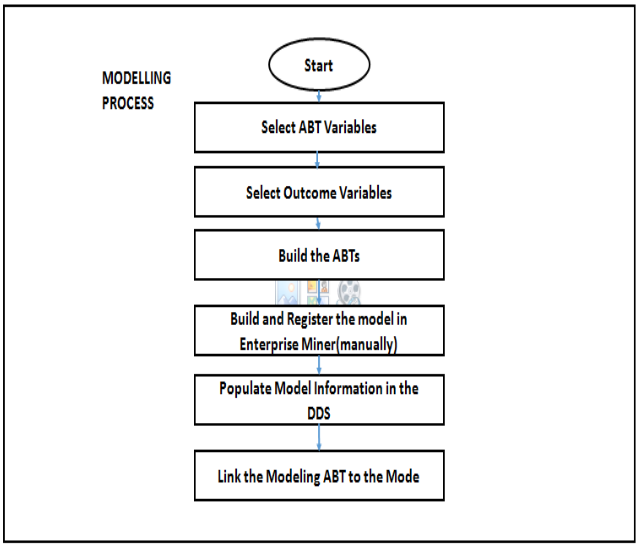

MODELLING PROCESS
Start
Select ABT Variables
Select Outcome Variables
Build the ABTs
Build and Register the model in Enterprise Miner(manually)
Populate Model Information in the DDS
Link the Modeling ABT to the Mode

Fig 3: Flowchart - working of Modeling ABT

The selected ABT variables are the required input variables and the outcome variables are the respective output variables for that requirement. After selection of input and output variables the model is built in SAS E-miner, the Model information is then populated into DDS. The next step is performed by the Scoring process.

Scoring ABT:
It is used for periodical scoring of data monthly and annually. It is more for an output operation

The significant scoring variables are selected, the scoring process is scheduled and the scores (periodic –Monthly, annually etc) are written back to DDS.

The ABTs have a major role in Input and process operations of required data from the specified problem definition.

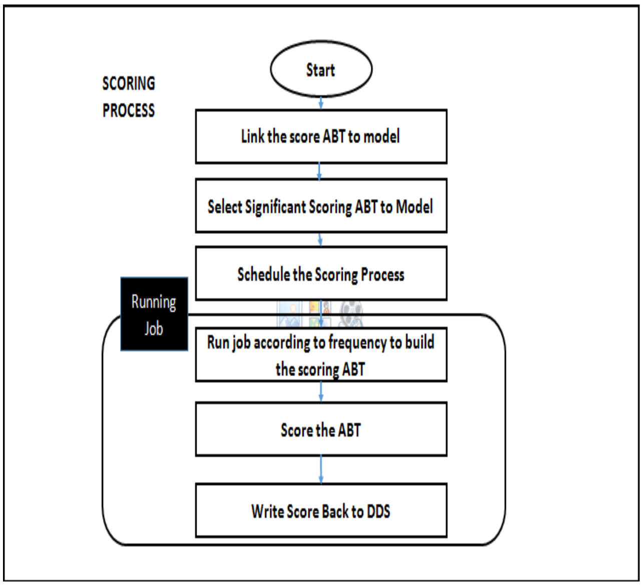

SCORING PROCESS
Start
Link the score ABT to model
Select Significant Scoring ABT to Model
Schedule the Scoring Process
Running Job
Run job according to frequency to build the scoring ABT
Score the ABT
Write Score Back to DDS

Figure 4: Flowchart - working of Scoring ABT

6. CREDIT SCORE MODEL

Credit Scoring is a process of building a statistical model to monitor the accuracy of the model and also the effect of the decisions taken based on the score obtained on key indicators [7]. Credit Scoring is used to generate Application Scorecards.

A credit scorecard is a numerical expression assigned to a customer. The scorecard development process has numerous data points (variables).One of them is target variable. Target variable would have the final model output (Yes/No) after following the entire process of Model Training, Testing and Model Validation.

Predictive Modeling concepts form an integral part in evaluating various risk parameters. In this paper, modeling of customer's Probability of Default (PD), a risk parameter which is required for credit risk analysis, is taken as an example. PD is computed using sample data using both parametric (linear and logistic) and non-parametric (Decision trees, Neural Networks etc.) models.

General form of linear regression is shown below. Where,

Yi = Probability of Default (0 to 1)

Xi – Xn = Predictor Variables (Gender, Salary, Region, Account type etc)

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1,\ldots,n,$$

A credit score is created using SAS E-miner. The data mining tool of SAS is SAS E-miner.[4][5]. Benefits of using SAS E-Miner in Credit Scoring:

Accuracy: Many credit scoring tools like CRMS are there, SAS E-Miner promises accurate results of high percentage.

User Friendly: Some tools such as SPSS already exist, however to develop a model using SPSS (and few of other existing tools) would take 2-3 months and can be done only by an expert. However SAS E-Miner is user friendly because a model can be developed within 10 days and with less expertise. The credit-scoring tab in it makes it easier to work on credit scoring.

Unbiased Sampling/Sampling Adjustment: In few credit scoring tools, the sample taken does not guarantee more than 75-85% of efficiency. The SAS E-miner, in most cases, guarantees 90-95% efficiency.

Scope for Non-Statisticians: Analytics is a domain where statistics knowledge background is mandatory to work on it. This was a big hindrance for non-statisticians to work in the field of analytics. The SAS E-miner tool is so user friendly as it performs many statistical operations within it (built-in) that even people from non-statistics can work in analytics.

SAS E-Miner Analytical Strengths: The analytical strengths of SAS E-miner include Pattern Discovery (to identify similar and dissimilar clusters, pattern matching), predictive modeling (to predict the future results, consequences etc.), and to perform Credit Scoring as to rate the customers.

Credit Scoring Models – Logistic Regression and Principal Component Analysis (PCA)- The two credit scoring models created, Logistic Regression and Principal Component Analysis are created for the same Dataset. **Logistic Regression (LR)**

Logistic Regression is applied to obtain a categorical value as output ('0' or '1'). It is a very widely used common credit scoring data model for predictive modeling.

Principal Component Analysis (PCA): Principal Component Analysis is a traditional multivariate statistical method commonly used to reduce the number of predictive variables. Principal Component Analysis is appropriate when one has to obtain measures on a number of observed variables and wish to develop a smaller number of artificial variables (called principal components) that will account for most of the variance in the observed variables. The principal components may then be used as a predictor or criterion variables in subsequent analyses

Advantages of Principal Component Analysis:

1. Numbers of variables are more in number while the data available (observations) are few in number.
2. In past applications, the internal variables of input Data set are used to calculate the results. In recent times the Credit Losses are estimated by using macro-economic variables to forecast predictions. There is a need for dimension reductionality.
3. It is used when the variables are highly correlated, and there is a need for large sample procedure. The PCA model is gaining more popularity and demands in current scenario .There are few scenarios where LR is preferred to PCA.

The value "Y "indicates that it is a better model which can be preferred out of the two models [10}.

Parameters used for model comparison:

The both credit scoring models (LR, PCA) created on same dataset can be compared .A model comparison is performed to conclude a better model amongst them. The parameters are:

1. Average Score Error (ASE): A model is compared to be better than others when its ASE value is less.
2. Misclassification Rate: The Misclassification Rate must be less for a better model.
3. Area under Curve (AUR): The Area under Curve is high for a better model.

Champion and Challenger models:

To evaluate the statistical reliability of the models, the champion model and challenger models enable you to compare proposed models using the same data .For any model and product category, it is recommended that one model be specified as the champion model. This model is known as the winner model. Other models are referred to as challenger models. There is only one champion model, but there are multiple challenger models. The champion model is routinely tested against challenger models.

Criteria to Evaluate the Model Performance:

The model created is tested for its performance by answering the below queries:

How efficient is the model in separating Good from the Bad customers.

Are the bad customers having low scores?

Are the Good customers having High scores?

7. SUMMARIZATIONS/ INFERENCES

The major issue in generating a model is Random Sampling Adjustments. It is a very challenging task and there is a possibility of misleading results and Graphs (Lift Chart, Gain Chart etc.).In the above models generated, accurate results are guaranteed. And, there is an option for more chances of equal proportion (for ex: 47: 50 – Good: Bad). The models built have high customer acceptance rate as it provides GUI data set creation, BASEL II Pooling and act as a compliant model of validation reports and OLAP business reports.

The prominent feature of the built Analytical Models are, it can calculate Probability of Default (PD), Loss Given Default (LGD) and Credit Conversion Factor (CCF).The credit scoring can be very easily integrated to the credit Risk Management for Banking.

8. ISSUES/DRAWBACKS/ LIMITATIONS

The data points available from the core banking system were not sufficient.

The data needs to be cleaned and enhanced for modeling.

Data sources are scattered across banks.

9. CONCLUSION

The credit scoring process satisfies Full BASEL II guidelines. It uses an integrated System, a data mart (DDS) to acquire data from source systems to ETL process, followed by selecting input data processing to generate reports. The credit scoring models are created for using logistic regression and principal component Analysis, better model is derived. The credit scoring process can be integrated very easily to Credit Risk Management System

10. FUTURE WORK

Data model in DDS to be enhanced to include other financial areas like Money Laundering, Fraud, Operational Risk etc.

Data model to be replicated to other Verticals like Insurance, Telecom etc

11. EXPERIMENTAL RESULTS/PROCESS

Data Sources

The data used for modeling is taken based on sample data obtained from Banks. The data variables selected are based on Business Logic considered for loans. The Business Logic involves Tenure of the Loan, Age parameter and others. The Business Logic/s involved in the creation of dataset is based on criteria known as DPD or Days Past Due.

DPD : Days Past Due is the time period given for the customers to repay their loans .If they are unable to repay with this specified period they are considered and categorized as Default.(in general, for most cases Default period in 90 days).

The parameters involved in the calculation of DPD and also the loan period for tenure of 3 years are considered and taken in the Dataset as the input variables. The input dataset consists of 500 observations, 22 input variable including one outcome variable .

Data Validity Checks

The data validation checks are performed to confirm the validity of the input data. It is performed by adopting various procedures:

Checking for Missing Values: The input dataset may not have the values

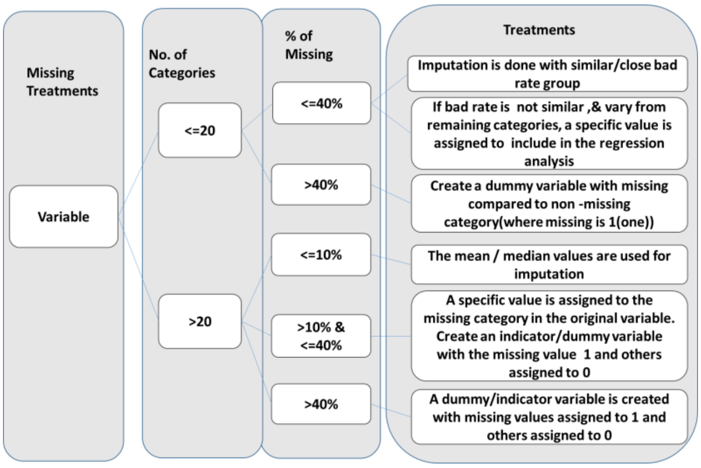
Anomaly Detection : Anomaly Detection or finding Outliers. Anomalies are those differently or unsuitable values in the Group.

Ex: age is 130 or -5 (minus five)

Validation of data on the missing check, quality of the data has been done using base sas queries.

The observations consisting of missing values need not be considered for creating a model. A good procedure that needs to be adopted is to impute Missing values before fitting into a model.

The input data set is transformed (on both dependent and independent variables) to make it better understandable and informative. It was observed that model created on transformed turns out to be better.



Imputation is done with similar/close bad rate group

If bad rate is not similar ,& vary from remaining categories, a specific value is assigned to include in the regression analysis

Create a dummy variable with missing compared to non -missing category(where missing is 1(one))

The mean / median values are used for imputation

A specific value is assigned to the missing category in the original variable. Create an indicator/dummy variable with the missing value 1 and others assigned to 0

A dummy/indicator variable is created with missing values assigned to 1 and others assigned to 0

Fig 5: Missing Value & output Treatment

The input dataset INPUT_PD has 22 variables included as per Business Logic parameters based on significance. The Dataset consists of 500 observations, total variables 22,one of the variable is target variable(DPD variable) .

SELECTION of INPUT VARIABLES based on TARGET variable value :

Selection of Input Variables:

Probabilities of Default (PD) models help validate the stability, performance, and calibration of models.

Modeling input requires the type of the variables from the input table in E-miner.

Basic Variables

These variables are categorized on the basis of the type of information that they store.

Behavioral

Variable that stores information about an account or a customer's behavior over a period of time. Typically, behavioral variables are observable and measurable characteristics or responses of subject. For example, the C_BAL_TOT_INQ_CNT_L12M variable shows the balance inquiry count, which shows how many times the customer inquired about the account balance during the past 12 months.

Time-Based

Variable that stores information about the last occurrence of a given activity in a defined time period. An example is a variable that shows the amount that a customer last withdrew through an ATM during the past six months. For example, LST_WDR_TR_CHN_L2M shows the channel that a customer last used to withdraw money during the past two months. LST_ATM_WDR_TR_AMT_3MB shows the amount of money a customer last withdrew through an ATM three months back.

Direct

A direct variable stores information about a particular attribute of a subject at a given point in time.

A direct variable typically captures demographic information about a customer, such as age, marital status, income, and city.

Derived Variables

These are variables that you can derive by using other variables, including basic and derived variables. Derived variables are categorized on the basis of how they are derived.

Data Validations:

The next step is to select the variables of input dataset based on target variable. Below is the list of selected input variables for Target variable Bad.

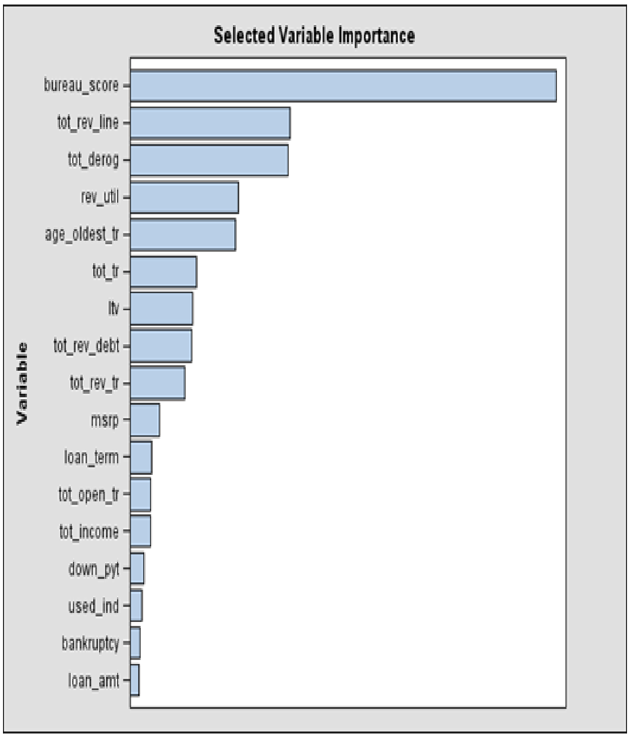
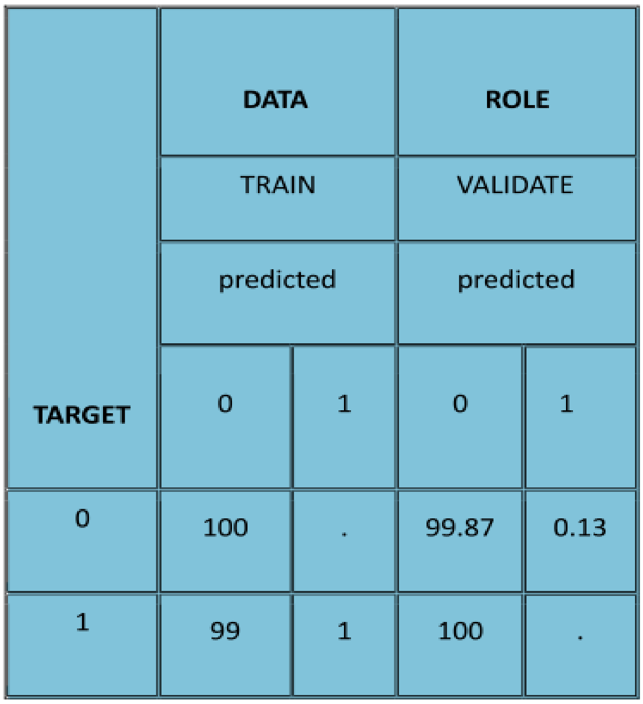


Fig 6: List of Selected Variable Importance

DATA PARTITION:

The next step is creation of TRAIN and VALIDATE datasets

Table 2: Classification Matrix



	DATA		ROLE	
	TRAIN		VALIDATE	
	predicted		predicted	
TARGET	0	1	0	1
0	100	.	99.87	0.13
1	99	1	100	.

After splitting the data into TRAIN and VALIDATE sets,

Interactive grouping process is applied on it.

INTERACTIVE GROUPING PROCESS:

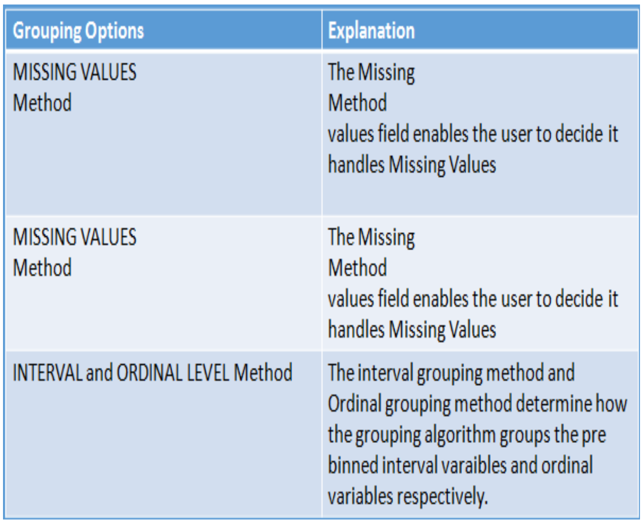
Grouping is done based on similar characteristics observed in the Data fields.

Interactive Grouping offers a number of advantages:

. Helps dealing with outliers

. Useful for understanding relationships

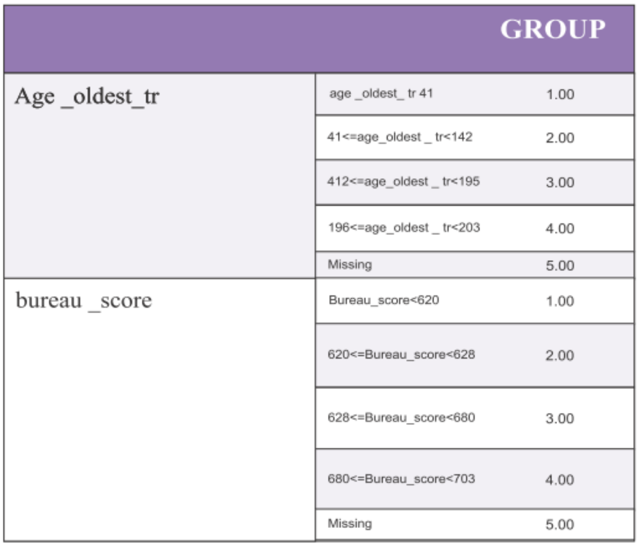
Table 3: Grouping Options/Procedures



Grouping Options	Explanation
MISSING VALUES Method	The Missing Method values field enables the user to decide it handles Missing Values
MISSING VALUES Method	The Missing Method values field enables the user to decide it handles Missing Values
INTERVAL and ORDINAL LEVEL Method	The interval grouping method and Ordinal grouping method determine how the grouping algorithm groups the pre binned interval varaibles and ordinal variables respectively.

The sample of the Grouping process is shown as below:

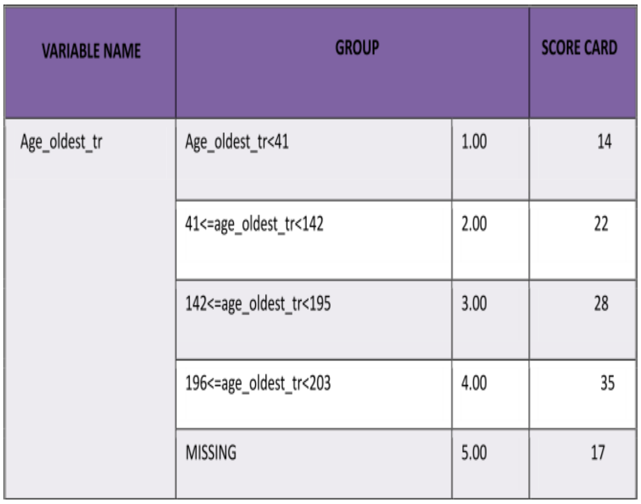
Table 4: Grouping Process – Sample



	GROUP	
Age _oldest_tr	age _oldest_ tr 41	1.00
	41<=age_oldest _ tr<142	2.00
	412<=age_oldest _ tr<195	3.00
	196<=age_oldest _ tr<203	4.00
	Missing	5.00
bureau _score	Bureau_score<620	1.00
	620<=Bureau_score<628	2.00
	628<=Bureau_score<680	3.00
	680<=Bureau_score<703	4.00
	Missing	5.00

Scaling the scorecard:
After the Grouping process the scorecard is scaled by applying appropriate statistical formulae.

Table 5: Scorecard- Sample



VARIABLE NAME	GROUP		SCORE CARD
Age_oldest_tr	Age_oldest_tr<41	1.00	14
	41<=age_oldest_tr<142	2.00	22
	142<=age_oldest_tr<195	3.00	28
	196<=age_oldest_tr<203	4.00	35
	MISSING	5.00	17

Creation of Model:
The Model is created using SAS E-Miner

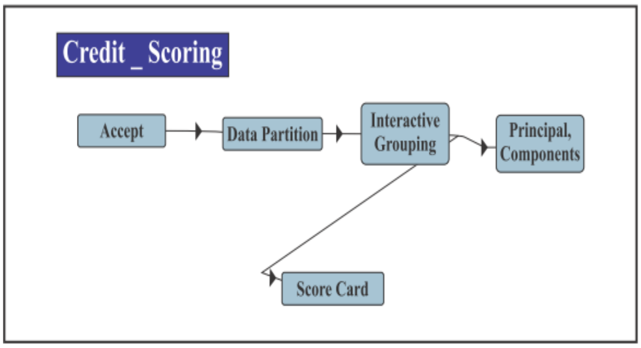


Fig 7: Creating a Model

Comparison of Models:
The models PCA, LR are compared to find the better amongst them

The following diagram depicts the comparison of two models for the same data set PCA and LR,

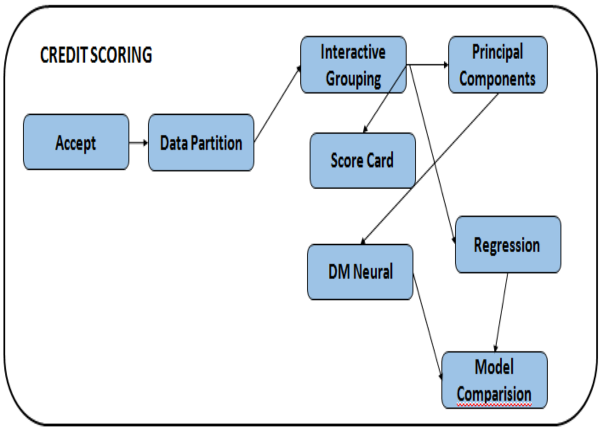
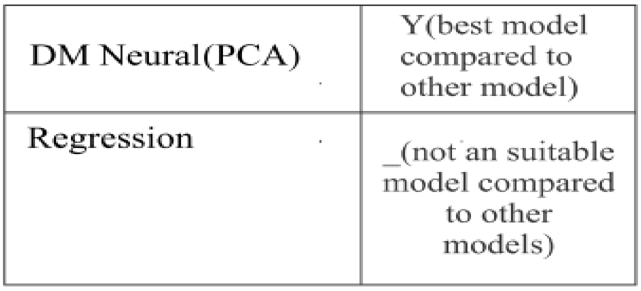


Fig 8: Model Comparision Diagram of PCA, LR

The partial output is as follows:

Table 7: Model comparison - Result Analysis



DM Neural(PCA)	Y(best model compared to other model)
Regression	_(not an suitable model compared to other models)

The value "Y "indicates that it is a better model which can be preferred out of the two models.

The scorecard is generated from interactive grouping node by applying various statistical procedures.

The final step includes generating two models on the same input dataset. The two models are compared and better model is chosen based on above mentioned model comparison parameters

12. ACKNOWLEDGMENTS

I thank Christ University for providing complete support on writing this paper. I exclusively thank Ms. Tulasi, Department of Computer Science, without whom this paper could not have been written who not only served as my supervisor but also encouraged and challenged me throughout my academic program. She guided me throughout the program, never accepting less than my best efforts. I also thank Info-drive Analytics for the complete support in terms of knowledge sharing, reviewing my views and helping me to shape this paper.

13. REFERENCES

[1] "Technical paper: Optimize ETL for the Banking DDS" http://support.sas.com/documentation/onlinedoc/dds/

[2] "SAS Detail Store for Banking: Implementation and User Guide- 2.0 to 4.8", http://support.sas.com/documentation/onlinedoc/dds/ddsadmin32.pdf

[3] "Pervasive SAS Techniques for Designing a Data warehouse for an Integrated Enterprise : An Approach towards Business process ",Ardhendu Tripathy et al,/(IJCSIT) International Journal of Computer Science and Information Technologies, vol.2(2), 2011,853-861Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[4] "Creating Interval Target Scorecards with Credit Scoring for SAS® Enterprise Miner™". Miguel Maldonado, Susan Haller, Wendy Czika, and Naeem Siddiqi SAS Institute Inc., Paper 094-2013

[5] "Building Loss Given Default Scorecard Using Weight of Evidence Bins in SAS® Enterprise Miner™", Anthony Van Berkel, Bank of Montreal and Naeem Siddiqi, SAS Institute Inc.", Paper 141-2012.

[6] "Risk and Compliance in Banking: Data Management Best Practices", white paper, www.sas.com/resources/ white paper. http://www.sas.com/resources/whitepaper/wp_65853.pdf

[7] "A Comparative Study of Data Mining Techniques for Credit Scoring in Banking",Shin-Chen Huang, Min-Yuh Day, Department of Information Management, Tamkang University, Taiwan DOI: 10.1109/IRI. 2013. 6642534 Publication Year: 2013 , Page(s): 684 – 691, **IEEE Conference** Publications

[8] "Default Predictors in Retail Credit Scoring : Evidence from Czech Banking Data",Evžen Kočenda , Martin Vojtek, Emerging Markets Finance & Trade, Vol. 47, No. 6 (November-December 2011), pp. 80-98, www.jstor.org/stable/41343442

[9] "Credit scoring for individuals", Maria DIMITRIU, Elena Alexandra AVRAMESCU, Razvan Constantin CARACOTA: Editura ASE: 2010 December : Economia : Seria Management, Vol 13, Iss 2, Pp 361-377 (2010): 1454-0320.

[10] Development of Credit Scoring Applications using SAS Enterprise Miner-User Guide