Credit Scorecards

Credit or application scorecards can be excellent tools for both lender and borrower to work out debt servicing capability of the borrower. For lenders, scorecards can help them assess the creditworthiness of the borrower and maintain a healthy portfolio – which will eventually influence the economy as a whole. Additionally to the borrower, they can provide valuable information such as 45% of people with her socio-economic background have struggled to keep up with the EMI commitment.

At a very high-level, credit scorecards have their roots in the classification problem in statistics & data mining. The classification problems present an extremely broad methodology/thought-process that has multiple business applications.

Including:
1. Classification problem and sampling
2. Variable selection and coarse classing
3. Predictive Models
4. Logistic regression and scorecards
5. Model validation
6. Application and business process integration

For more details about score cards read the book which was downloaded at another PC

Credit Scorecards – Classification Problem

In the case of credit scorecards, the problem statement is to distinguish analytically between the good and bad borrowers. Hence, the first task is to define a good and a bad borrower. For most loan products, good and bad credit is defined in the following way

1. Good loan: never or once missed on the EMI payment
2. Bad loan: ever missed 3 consecutive EMIs in a row (i.e. 90 days-past-due)

Additionally, for tagging someone good or bad, you need to observe his or her behavior for a significant length of time. This length of time varies from product to product based on the tenor of the loan. For home loans, with a tenor of 20 years, 2-3 years is a reasonable observation period.

Sampling Strategy for Credit Scorecards

Black swan problem.
Example: You may need to visit every single planet to rule our possibility of an intelligent form of life.

For credit scorecard development, the accepted rule of thumb for sample size is at least 1000 records of both good and bad loans.

Variables Selection

$$GDP = C + I + G + (Ex - Im)$$

C = *collective spending by consumers*

I = *cumulative investment by businesses*

G = *total spending by government*

$(Ex - Im)$ = *net exports (exports − imports)*

Gross Domestic Product

As per the formulae the above mentioned parameters are the best fit to calculate GWP. But there could be better formulae which may include more parameters to calculate GWP as India is too diversified.
The idea is to select the right variables to build your model!

variables selection process is performed through statistical significance – a reasonably automated process.

Data Collection:
Application forms are a major source of data collection regarding the borrower. However, nobody wants to fill a lengthy form hence an optimal size of the form ensures accurate information provided by the borrower. The idea is to select the right variable and ensure accurate measurement.

There are several aspects regarding variables but I will mention just one of them here (coarse classing).

Coarse Classing as a technique during model development.

Quite a few academicians & practitioners for a good reason believe that coarse classing results in loss of information. However, in my opinion, coarse classing has the following advantage over using raw measurement for a variable.

1. It reduces random noise that exists in raw variables – similar to averaging and yes, you lose some information here.

2. It handles extreme events – on two extremes of a variable – much better where you have thin data.

3. It handles the non-linear relationship between dependent and independent variables without a lot of effort of variable transformation from the analyst.

PCA could be better, or AIC or variable importance plot

	Original Data				Coarse Classes					
Age Group	Total Number of loans	Number of Bad loans	Number of Good Loans	% Bad loans	Age Group	Total Number of loans	Number of Bad loans	Number of Good Loans	% Bad loans	Name of Coarse Groups
21-24	310	14	296	4.5%						
24-27	511	20	491	3.9%	**21-30**	**4821**	**206**	**4615**	**4.3%**	G_1
27-30	4000	172	3828	4.3%						
30-33	4568	169	4399	3.7%						
33-36	5698	188	5510	3.3%	**30-36**	**10266**	**357**	**9909**	**3.5%**	G_2
36-39	8209	197	8012	2.4%						
39-42	8117	211	7906	2.6%						
42-45	9000	216	8784	2.4%	**36-48**	**32926**	**776**	**32150**	**2.4%**	G_3
45-48	7600	152	7448	2.0%						
48-51	6000	84	5916	1.4%						
51-54	4000	64	3936	1.6%						
54-57	2000	26	1974	1.3%	**48-60**	**12788**	**183**	**12605**	**1.4%**	G_4
57-60	788	9	779	1.1%						

Table 1 – Coarse Class

MODELLING:

A model is defined as a simplified representation of reality.

Data warehouse, Business Intelligence and Advanced Analytics

Analytics has received a massive boost because of the emergence of information technology. We are living in the era of big data. A plethora of data collected at every stage of the business process had created a need to extract knowledge out of the information. This overall process has three aspects to it

1. **Data warehouse or data marts:** transactional data is extracted-transformed and loaded (ETL) into a data model / schema for the purpose of analysis

2. **Business Intelligence or dashboards:** “as is” business reports

3. **Predictive Analytics or Advanced Analytics:** high-end statistical and data mining exercise

As the quantum of data is exponentially increasing, **Hadoop and big data** technologies are replacing the data warehouses.

Credit Scoring Models

Credit scorecards are models to predict the probability of a borrower default on his/her loan. The following is a simplified version of credit score with three variables

Credit Score = Age + Loan to Value Ratio (LTV) + Installment (EMI) to Income Ratio (IIR)

Points Table					
Age of the applicant (in years)		**Loan to Value Ratio (LTV)**		**Instalment to income ratio(IIR)**	
Below 32 years	10	0 to 50	100	Below 20	140
32 to 50 year	50	50 to 80	50	20 to 50	75
Above 50 years	20	80 to 100	10	Above 50	5

Score Points wise Risk					
Below 100	High Risk	100 to 180	Medium Risk	Above 180	Low Risk

A 28-year-old man with the LTV of 75 and the IIR of 60 will have the score of 10+50+5 =65 and hence is a high credit risk.

There are several statistical & data mining techniques that could help us achieve our object such as

1. Decision tree
2. Neural Networks
3. Support Vector Machines
4. Probit Regression
5. Linear discriminant analysis
6. Logistic Regression

Logistic regression is the most commonly used technique for the purpose. We will explore more about logistic regression in the next article.

Applications of logistic regression:
• Good and Bad borrowers
• Fraud and genuine cases
• Buyers and non-buyers

Out of sample test: where we have divided our sample into the training and the test sample. The first level of testing happens on the holdout or test sample. The test sample needs to perform as well as the training sample.

Out of time sample test: since the model was built on a sample of the portfolio with reasonable vintage, the analyst would like to test the performance of a more recent portfolio.

On field test: this is where the test of the pudding is; the analyst needs to be completely aware of any credit policy changes that the bank has gone through since the scorecard is developed and more importantly, the impact the changes will have on the scorecard.

Performance Tests for Model Validation

There are several ways to test the performance of the scorecard such as c**onfusion matrix, KS statistics, Gini and area under ROC curve (AUROC)** etc. The KS statistics is a widely used metric in scorecards development. However, I personally prefer the AUROC to the others. I must add the Gini is a variant of the AUROC. The reason for my liking of the AUROC could be my formal training in Physics and engineering. I think it is a more holistic measure and lets the analyst visually analyze the model performance. I prefer graph and visual statistics any day to raw numbers.

BANKING ANALYTICS

Financial Ratios

When corporate analysts try to analyze financials of a company they often work with several financial ratios. Working with ratios has a definite advantage over working with plain vanilla variables. Combined variables often convey much higher information. Seasoned analysts understand this really well. Moreover, variable creation is a creative exercise that requires sound domain knowledge. For credit analysis, the ratio of the sum of obligations to income is highly informative since this provides an insight about percentage disposable income for the borrower.

Example Links:
https://github.com/brunokatekawa/credit_risk/blob/master/Credit_Risk.ipynb