CREDIT RISK ANALYZER WITH AI

-By Vaibhav Vats

Introduction

What is Credit Analysis?

Credit analysis is a type of analysis an investor or bond portfolio manager performs on companies or other debt issuing entities to measure the entity's ability to meet its debt obligations. The credit analysis seeks to identify the appropriate level of default risk associated with investing in that particular entity.

What type of information is critical?

Some types of loans require more thorough analysis than others. Larger, long-term loans for fixed assets require more thorough analysis than short-term working capital loans. For individual loans, loan analysis and follow-up visits provide most of the guarantee for the institution and thus the analysis is necessarily more extensive. Group loans transfer most of this responsibility to the clients and therefore do not require detailed analysis.

Credit Risk Predictive Modelling and Credit Risk Predictive By Machine Learning

If past is any guide for predicting future events, credit risk prediction by Machine Learning is an excellent technique for credit risk management. Prediction models are developed from past historical records of credit loans, containing financial, demographic, psychographic, geographic information, etc. From the past credit information, predictive models can learn patterns of different credit default/delinquency ratios, and can be used to predict risk levels of future credit loans. It is important to note that statistical process requires a substantially large number of past historical records (or customer loans) containing useful information. Useful information is something that can be a factor that differentially affects credit default/delinquency ratios.

Objective

To ensure that loans are made on appropriate terms to clients who can and will pay them back. What analysis is needed and what is the most efficient approach to fulfil that need is primarily determined by the type and nature of the loan.

• To place good and appropriate loans - can the loan generate income for repayment and will the client repay

• Determine eligibility of the applicant - is he/she eligible according to the the program criteria

• Training needs and skills - to assess the training needs and develop the financial management skills level of the client. (This is the basic principal of programs that integrate their credit and training methodologies.)

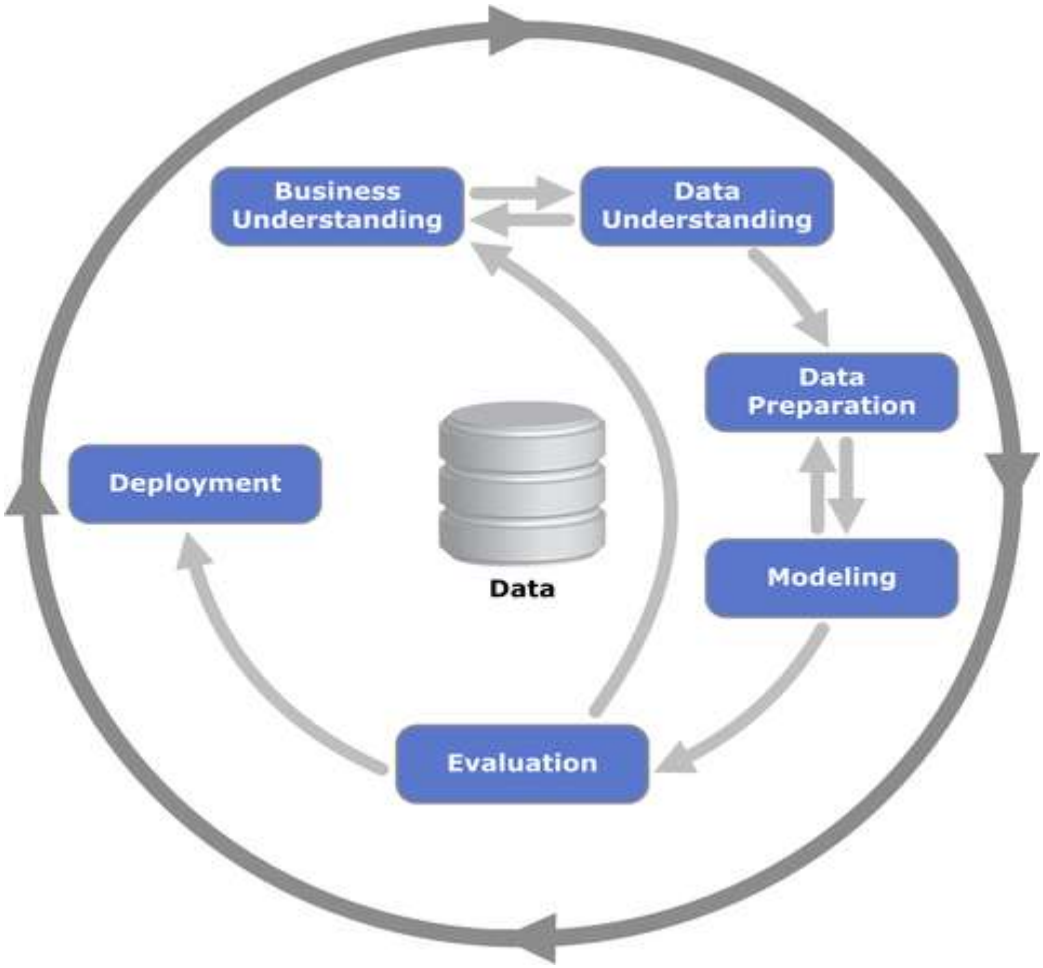
• Program Indicators - loan analysis may also be used to generate the indicators that will be used to evaluate the impact of the loan.

Dataset

age	gender	education	occupation	organization_type	seniority	annual_income	disposable_income	house_type	vehicle_type	marital_status	no_card	default
19	Male	Graduate	Professional	None	None	186319	21625	Family	None	Married	0	1
18	Male	Under Graduate	Professional	None	None	277022	20442	Rented	None	Married	0	1
29	Male	Under Graduate	Salaried	None	Entry	348676	24404	Rented	None	Married	1	1
18	Male	Graduate	Student	None	None	165041	2533	Rented	None	Married	0	1
26	Male	Post Graduate	Salaried	None	Mid-level 1	348745	19321	Rented	None	Married	1	1
26	Female	Other	Student	None	None	404972	22861	Family	None	Single	0	1
28	Male	Under Graduate	Student	None	None	231185	20464	Family	None	Married	0	1
24	Female	Under Graduate	Salaried	None	Entry	102554	42159	Family	None	Married	1	1
26	Female	Under Graduate	Salaried	None	Junior	226786	19817	Family	None	Single	0	1
26	Male	Graduate	Salaried	None	Mid-level 1	250424	5271	Family	Two Wheeler	Married	1	1
21	Male	Graduate	Professional	None	None	154970	1916	Family	None	Married	2	1
20	Male	Graduate	Business	None	None	160487	14949	Family	None	Married	2	1
19	Male	Under Graduate	Salaried	None	Junior	311863	6142	Rented	None	Single	1	1
22	Male	Graduate	Professional	None	None	307890	41440	Family	Two Wheeler	Single	0	1
28	Male	Other	Professional	None	None	165164	16114	Rented	Two Wheeler	Other	1	1
26	Male	Under Graduate	Student	None	None	101988	2625	Rented	None	Other	1	1
27	Female	Other	Professional	None	None	322711	3016	Rented	None	Single	1	1
25	Female	Graduate	Business	None	None	128971	13305	Family	Two Wheeler	Single	0	1
29	Female	Under Graduate	Professional	None	None	165658	8494	Rented	Two Wheeler	Single	1	1
28	Male	Graduate	Professional	None	None	425478	10286	Family	Two Wheeler	Single	1	1
29	Male	Post Graduate	Professional	None	None	345982	20479	Rented	None	Single	2	1
21	Male	Graduate	Salaried	Tier 3	Entry	133975	8762	Family	None	Single	0	1
19	Female	Other	Salaried	None	Entry	269009	2865	Rented	None	Married	1	1
23	Female	Other	Business	None	None	273606	11495	Family	None	Married	1	1
24	Male	Graduate	Salaried	Tier 3	Entry	276982	18540	Family	None	Married	0	1
23	Male	Under Graduate	Student	None	None	235798	39372	Rented	None	Married	0	1
18	Male	Post Graduate	Salaried	None	Mid-level 1	422071	25861	Family	None	Single	0	1
21	Male	Post Graduate	Salaried	None	Mid-level 1	473825	35966	Rented	None	Other	0	1
26	Female	Graduate	Student	None	None	63362	19780	Rented	None	Married	0	1
22	Female	Post Graduate	Professional	None	None	54005	8829	Rented	Two Wheeler	Married	0	1

Methodology and Techniques Used

Cross-industry standard process for data mining



CRISP-DM stands for cross-industry process for data mining. The CRISP-DM methodology provides a structured approach to planning a data mining project. It is a robust and well-proven methodology. We do not claim any ownership over it. We did not invent it. We are however evangelists of its powerful practicality, its flexibility and its usefulness when using analytics to solve thorny business issues.

Python Modules Used

- Scikit-learn (sklearn)
  - DecisionTreeClassifier
  - KFold
  - train_test_split
  - accuracy_score
- Matplotlib
- Pandas
- NumPy

Decision Tree

A **decision tree** is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

A decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

There are three commonly used impurity measures used in binary decision trees: **Entropy**, **Gini index**, and **Classification Error**.

Entropy (a way to measure impurity):

Entropy=−Sum (p * log2p)

Gini index (a criterion to minimize the probability of misclassification):

Gini=1−Sum (p * p)

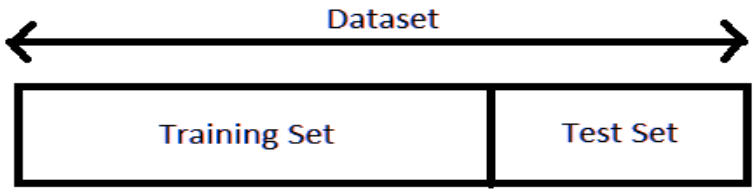
Classification Error:

ClassificationError=1−max(p)

where p is the probability of classes.

Train Test Split

As we work with datasets, a **machine learning algorithm** works in two stages. We usually split the data around 20%-80% between testing and training stages. Under supervised learning, we split a dataset into a training data and test data in Python ML.



The following section will split the dataset randomly into two groups, training dataset and test dataset. We will use 70% data as training data and remaining 30% as test data.

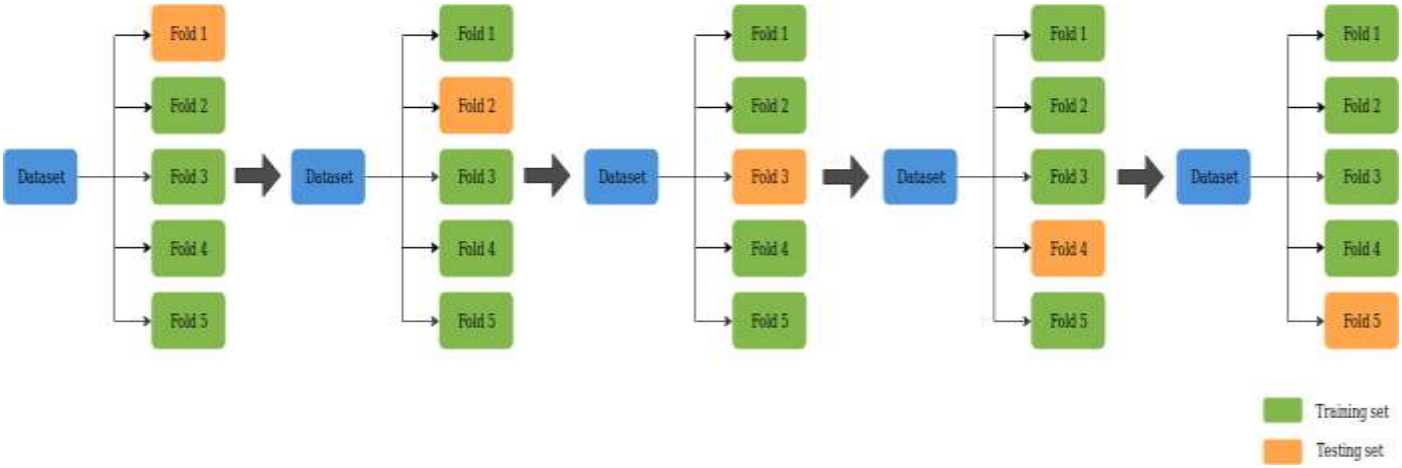
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test= train_test_split (X, Y, test_size=0.3, random_state=1234)

K-fold Cross Validation

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.



In this project value of k is 4.

Train Decision Tree Model with Gini and Entropy Criterion-

- K-Fold (4-Fold) Cross validation to attain high accuracy
- Predict data for every fold

Picking average of scores

Model Summary

- 84.46% (Gini Criterion with K-Fold)
- 84.29% (Entropy Criterion with K-Fold)
- 84.08% (Gini Criterion with train_test_split)
- 83.92% (Entropy Criterion with train_test_split)