CREDIT RISK ANALYZER

* **Amit Kumar**

Introduction

**Credit analysis** is the method by which one calculates the creditworthiness of a business or organization or any particular entity. In other words, it is the evaluation of the ability of a company or an individual to honor its financial obligations.

A **credit card** is a cashless transaction system that is linked with an account in a bank that issued the credit card. This account is called credit card account. In its physical form it is a plastic card that we need to swipe at merchant PoS machine to complete a transaction. We can also use the credit card details online to purchase things online.

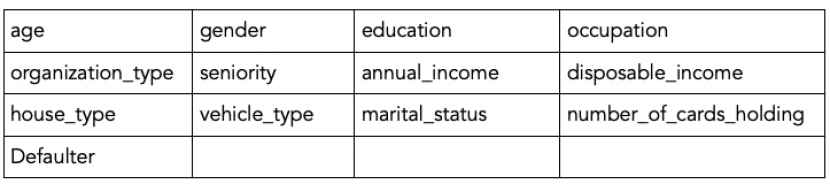
But how does this happen? The account linked with the credit card, has a given credit limit. When we purchase something using the credit card the money gets deducted from the credit limit and paid to the merchant. We have to return the money to the account as per the billing cycle. Now, this credit limit is not the amount that we have paid in advance to the bank and kept for future usages through the credit card. This is essentially an amount that the bank has assigned to us, in our credit card account, as some sort of approved loan amount. From where we can borrow money through the credit card at the merchant PoS machine or online, and return the money as per our approved repayment plan. This credit limit is a risk that bank is taking on you. Why? Because, if you use the money and do not return that to the bank, the bank loses the entire amount. This is a **credit risk**.

Banks need to protect their interest before it can take risk on you and issue credit card to you. Banks use their previous credit card holder’s records for understanding the patterns of the card holders. It is a lot more complex process to predict whether a person who they do not know at personal level, will be a defaulter or not. Banks, along with the data from their own records, also use CIBIL data. CIBIL plays an important role here as it collects data from all your loan types, across financial institutes, over a period of time. As it is defined, “CIBIL Score is a three-digit numeric summary of your credit history. The score is derived using the credit history found in the CIBIL Report (also known as CIR i.e. Credit Information Report). A CIR is an individual’s credit payment history across loan types and credit institutions over a period of time. A CIR does not contain details of your savings, investments or fixed deposits.” Higher the score better it is. The banks also consider market conditions, their current risk appetite etc. before they issue a credit card to you.

Based on all this data, banks want to develop a pattern that will tell them who are likely to be a defaulter and who are not. As an example, in its simplest form just for our understanding, let’s say, people who are salaried and own a house have lower number of default cases as compared to self-employed people who are on rent. Or let’s say people at the higher income group have lesser cases of defaulting as compared to people belonging to the lower income group. If that are what the banks identify as patterns of defaulting, they would like to test their formula with some data first. If the data support their formula, the banks would use this formula as an identifier of suspected defaulter, and issue credit cards based on that. This formula or rule or technique is credit risk analyzer. Banks collect applicant’s data from their credit card application forms. They also access the applicant’s CIBIL score. They use all the relevant information they need in their credit risk analyzer. All data are fed into the analyzer and based on the rules set in the analyzer an applicant’s application is accepted or rejected.

Dataset

We have the data for the following variables.

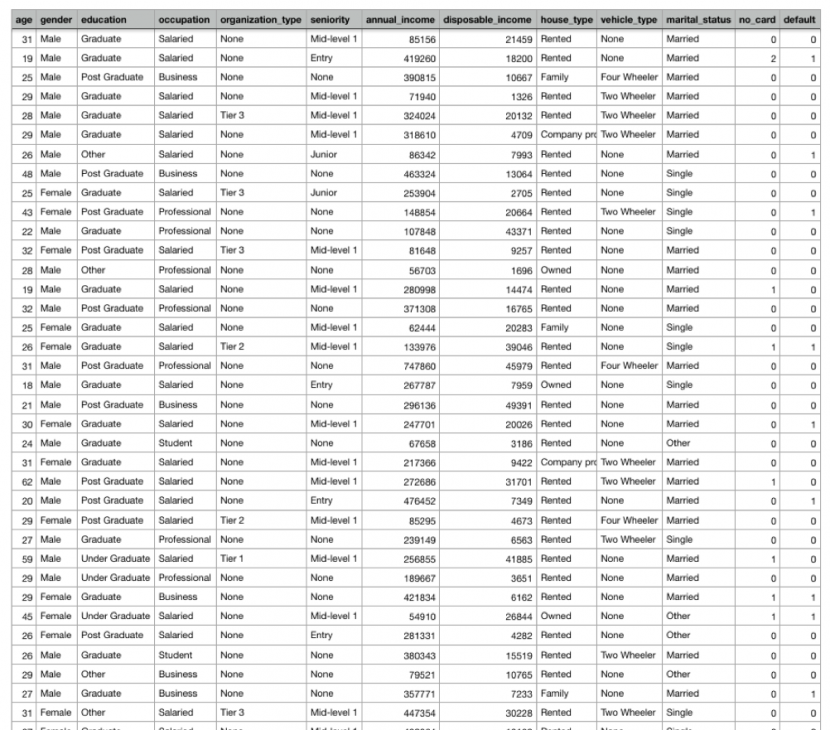


Here age, gender are the age and gender of the card holder. Education is the last acquired educational qualification of the card holder. Occupation can be salaried, or self-employed or business etc. Organization type can be tire 1, 2, 3 etc. Seniority denotes at which career level the card holder is in. Annual income is the gross annual income of the card holder. Disposable income is annual income - recurring expenses. Recurring expenses can be EMIs etc. House type is owned or rented or company provided etc. Vehicle type is 4-wheeler or two-wheeler or none. Marital status is of the card holder. This data also has the information of the number of other credit cards that the card holder already holds. And at the end of each row, we have a defaulter indicator indicating whether the card holder was a defaulter or not. It is 1 if the card holder was a defaulter, 0 otherwise.

The data contains exhaustive information about card holders. Patterns based on these variables will indicate the likelihood of a new applicant to be a defaulter.

First, we will build a decision tree credit risk analyzer, based on a portion of the data. Second, we will test the credit risk analyzer based on the remaining data. If we have found an analyzer that is tested ok, we will use the analyzer to classify new applicants as per their likelihood of being a defaulter.

The data in the .csv format is added to your dashboard. A snapshot of data is as follows.

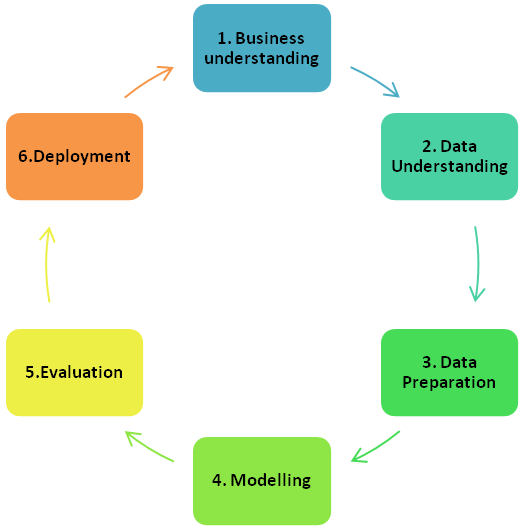


Methodology and Techniques Used

CRISP-DM

**CRISP-DM methodology** stands for cross-industry process for data mining. This was developed in 1996 under ESPRIT initiative. This methodology provides a structured approach for planning and managing a data science projects. It is a robust and well-proven methodology. Business analysts and data scientists mostly prefer this model as it is easily adaptable.

The CRISP-DM model is shown as follows:



CRISP-DM is an idealised sequence of events as shown above. In practice many of the tasks can be performed in a different order and it will often be necessary to backtrack to previous tasks and repeat certain actions.

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)

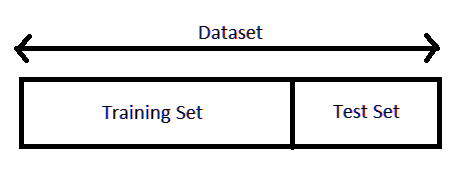
**Classification Error**:

Classification Error=1−max(p)

where p is the probability of classes.

Train Test Split

As we work with datasets, a [**machine learning algorithm**](https://data-flair.training/blogs/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.

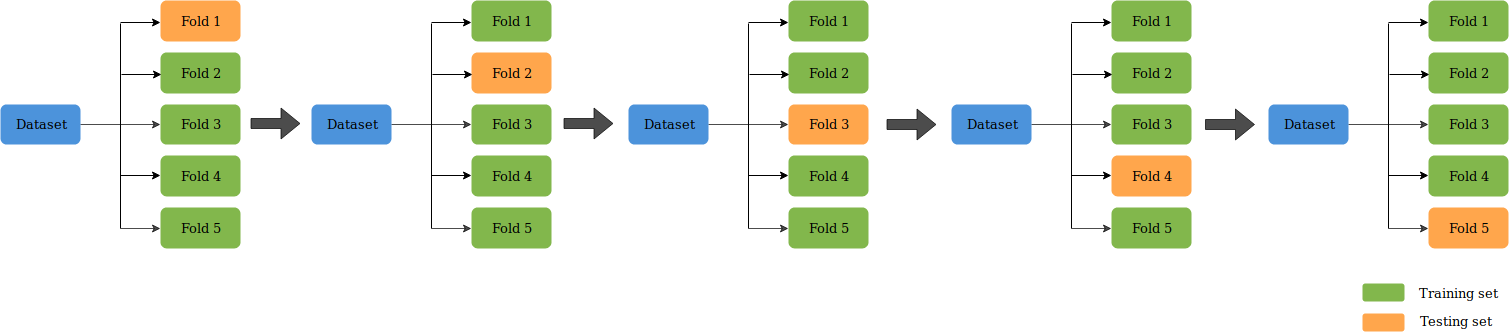
*from sklearn.model\_selection import train\_test\_split*

*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

* 84.6% (Gini Criterion with K-Fold)
* 84.36% (Entropy Criterion with K-Fold)
* 84.58% (Gini Criterion with train\_test\_split)
* 83.55% (Entropy Criterion with train\_test\_split)