8/24/2020

By- Himanshu Mittal

Student

(Manipal university Jaipur)

Credit Risk Analyzer

**Introduction**

**Credit Analysis**

**Credit analysis** is the method by which one calculates the [creditworthiness](https://en.wikipedia.org/wiki/Creditworthiness) of a business or organization. In other words, it is the evaluation of the ability of a company to honour its financial obligations. The audited financial statements of a large company might be analysed when it issues or has issued [bonds](https://en.wikipedia.org/wiki/Bond_(finance)). Or, a [bank](https://en.wikipedia.org/wiki/Bank) may analyse the financial statements of a small business before making or renewing a commercial loan. The term refers to either case, whether the business is large or small.

The objective of credit analysis is to look at both the borrower and the lending facility being proposed and to assign a risk rating. The risk rating is derived by estimating the probability of default by the borrower at a given confidence level over the life of the facility, and by estimating the amount of loss that the lender would suffer in the event of default.

Credit analysis involves a wide variety of financial analysis techniques, including [ratio](https://en.wikipedia.org/wiki/Ratio) and trend analysis as well as the creation of projections and a detailed analysis of cash flows. Credit analysis also includes an examination of [collateral](https://en.wikipedia.org/wiki/Collateral_(finance)) and other sources of repayment as well as credit history and management ability. Analysts attempt to predict the probability that a borrower will default on its debts, and also the severity of losses in the event of default.

**Credit Scoring systems**

In recent decades, a number of objectives, quantitative systems for scoring credits have been developed. In [univariate](https://en.wikipedia.org/wiki/Univariate) (one variable) [accounting](https://en.wikipedia.org/wiki/Accounting)-based credit-scoring systems, the credit analyst compares various key accounting ratios of potential borrowers with industry or group norms and trends in these variables.

Today, [Standard & Poor's](https://en.wikipedia.org/wiki/Standard_%26_Poor%27s), [Moody's](https://en.wikipedia.org/wiki/Moody%27s), and [Risk Management Association](https://en.wikipedia.org/w/index.php?title=Risk_Management_Association&action=edit&redlink=1) can all provide banks with industry ratios. The univariate approach enables an analyst starting an inquiry to determine whether a particular ratio for a potential borrower differs markedly from the norm for its industry. In reality, however, the unsatisfactory level of one ratio is frequently mitigated by the strength of some other measure. A firm, for example, may have a poor [profitability ratio](https://en.wikipedia.org/w/index.php?title=Profitability_ratio&action=edit&redlink=1) but an above-average [liquidity ratio](https://en.wikipedia.org/wiki/Quick_Ratio). One limitation of the univariate approach is the difficulty of making trade-offs between such weak and strong ratios. Of course, a good credit analyst can make these adjustments. However, some univariate measures – such as the specific industry group, public versus private company, and region – are categorical rather than ratio-level values. It is more difficult to make judgments about variables of this type.

Although univariate models are still in use today in many banks, most academics and an increasing number of practitioners seem to disapprove of ratio analysis as a means of assessing the performance of a business enterprise. Many respected theorists downgrade the arbitrary rules of thumb (such as company ratio comparisons) that are widely used by practitioners and favor instead the application of more rigorous statistical techniques.

**Credit Risk Predictive modelling and Credit risk predictive by machine learning**

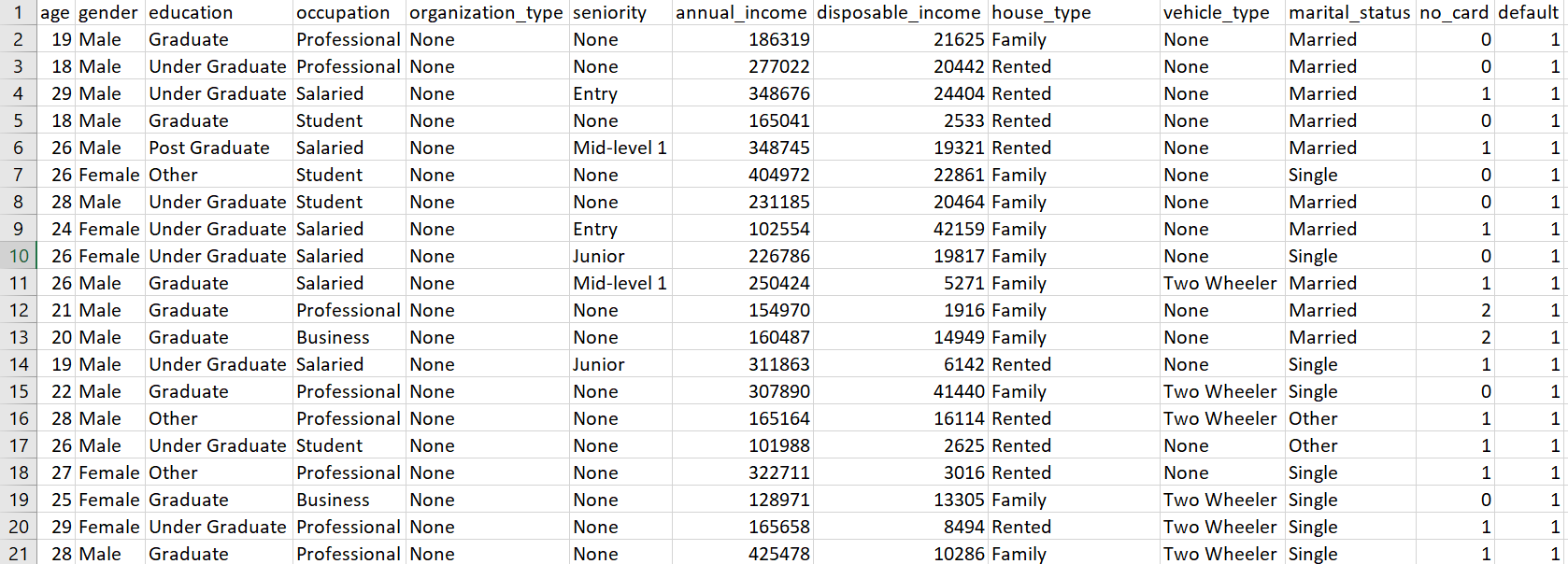
Credit risk predictions can be achieved using various machine learning and deep learning techniques like Decision Tree, Logistic regression, Artificial neural network etc. Prediction models are developed from past historical records of credit loans, containing financial, demographic, geographic location, education details, marital status

Predictive models from learn patterns from different credit default 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

**Objective**

To ensure that loans are made on appropriate terms to clients who can and will pay them back and 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.

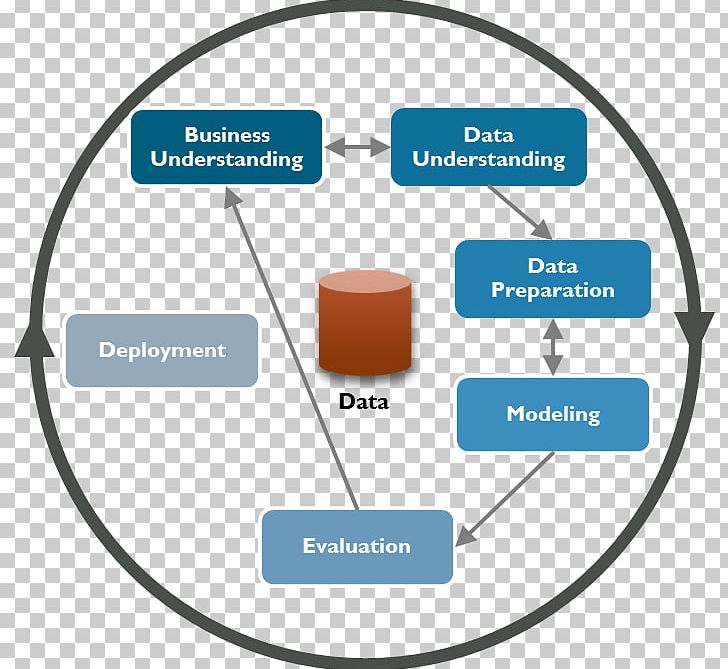
**Dataset**



The Dataset contains the information of loan applicants including their education, age, gender, occupation, annual income, marital status etc.

**Methodology and technique used**

**Cross-industry Standard Process for Data Mining**



Cross-industry standard process for data mining, known as CRISP-DM, is an [open standard](https://en.wikipedia.org/wiki/Open_standard) process model that describes common approaches used by [data mining](https://en.wikipedia.org/wiki/Data_mining) experts. It is the most widely-used [analytics](https://en.wikipedia.org/wiki/Analytics) model.

CRISP-DM breaks the process of [data mining](https://en.wikipedia.org/wiki/Data_mining) into six major phases

* Business Understanding
* Data Understanding
* Data Preparation
* Modelling
* Evaluation
* Deployment

**Python Modules used:**

* Scikit-learn(sklearn)
  + Decision TreeClassifier
  + Train\_test\_split
  + Accuracy\_score
  + Confusion matrix
  + Roc\_curve
* Matplotlib
* Pandas
* NumPy

**Decision Tree Classification**

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, each representing values for the attribute tested. Leaf node represents a decision on the numerical target. The topmost decision node in a tree, which corresponds to the best predictor, called root node. Decision trees can handle both categorical and numerical data.

**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.

**Implementation**

1. **Pre-processing**

X = Credit\_data.drop([‘default’], axis = 1)

y = Credit\_data[‘default’]

1. **Train\_test\_split**

X\_train, X\_test, y\_train, y\_test = train\_test\_split (X, y, test\_size = 0.3, random\_state = 1)

1. **Machine learning model(Entropy)**

credit\_tree = DecisionTreeClassifier(criterion = "entropy", max\_depth = 5)

credit\_tree.fit(x\_train,y\_train)

1. **Prediction and accuracy**

pred\_tree = credit\_tree.predict(x\_test)

print("Accuracy:", accuracy\_score(y\_test, pred\_tree)\*100)

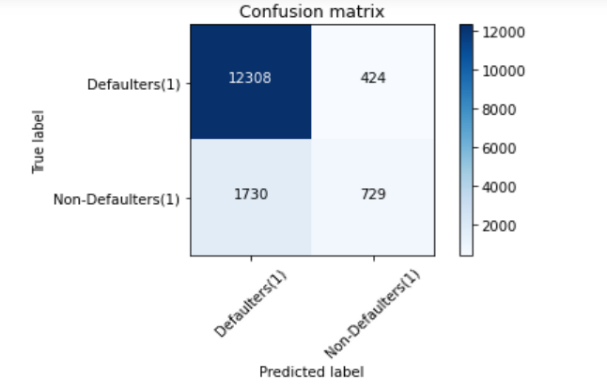
1. **Confusion matrix**

dt = confusion\_matrix(y\_test, pred\_tree)

np.set\_printoptions(precision=2)

plot.figure()

plot\_confusion\_matrix(dt, classes = ['Defaulters(1)', 'Non-Defaulters(1)'], normalize=False, title='Confusion matrix')



1. **ROC Curve**

plot.title('Receiver Operating Characteristic')

plot.plot(fpr,tpr,'b', label = 'AUC = %0.2f' % roc\_auc)

plot.legend(loc = 'lower right')

plot.plot([0, 1], [0, 1],'r--')

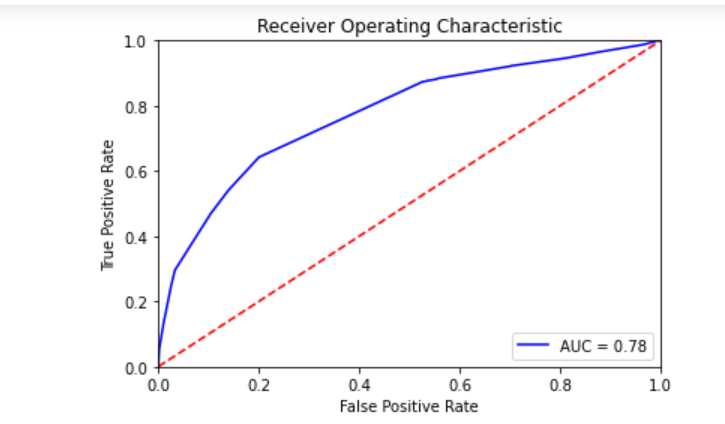
plot.xlim([0, 1])

plot.ylim([0, 1])

plot.ylabel('True Positive Rate')

plot.xlabel('False Positive Rate')

plot.show()



**Result**

* Accuracy with entropy criterion: 85.81
* Accuracy with Gini criterion: 85.82
* Area under curve: 0.78