**Credit Scoring**

* Credit scoring is a statistical analysis performed by lenders to determine the creditworthiness of a person or small, owner operated business.
* Creditworthiness is how a lender determines that you are worth to apply and get loans. It is determined by several factors such as your debt credit report, credit score and repayment history.

**Credit Report**

* Credit report is detailed information about credit history.
* It includes personal information such as names, addresses, marital status and employment history.
* It also covers number and type of bank accounts and all their status; credit limits and date accounts were opened.
* It will also show information regarding any missed payments and public records regarding financial transactions and whether you have filed for bankruptcy.
* One last thing is that it includes all the entities that have asked to see credit report because of events such as applying for a personal loan.

**Credit Reporting Agencies**

* Credit reporting agencies (CRAs) are the ones that collects and maintains the credit reports.
* They collect the historical credit information on individuals and businesses from financial institutions such as banks and credit card companies.
* CRAs are the ones who generates the credit scores based on the credit reports they generated.
* Egypt’s credit reporting agency is I-score owned by an association of financial institutions that includes Central Bank of Egypt (CBE), banks and microfinance institutions.

**Credit Score**

* Credit score is based a number that reflects the consumer creditworthiness. So, it helps lender decide whether to extend or deny credit.
* **Credit scores can impact a person’s financial transactions such as the ability to borrow money for credit cards, mortgages and loans.**
* Many factors contribute to credit scores including payment interest, previous and current loans and debts, the type of debts, length of credit history, payment history.
* **The most famous and used credit scoring system is FICO system which assigns a numerical representation of creditworthiness that ranges from 300 to 850 where the higher the number, the higher the individuals credit score.**
* Getting a low credit score doesn’t always mean that he/she can’t apply for a loan. It just makes it more likely to pay more to borrow money than someone with a higher credit score. It also effects the repayment period.
* **The reason is that the lender is taking a risk by approving the debt and a person with a low credit score could not have the guarantees needed for reassurance and proving that this person is going to pay the payments on time and won’t miss on payments. So, to compensate that risk, they receive the loan with a higher interest are and with shorter repayment period.**
* Credit scoring systems establishes a guideline for the lenders to determine which level is acceptable and how much to charge in interest.

**Statistical Techniques**

**Logistic Regression**

* **Predictive regression analysis that is used to describe relation between one dependent variable and one or more independent variables.**
* The independent variables could be continuous or categorical but the output is dichotomous(binary).
* **Logistic regression is used to calculate the probability of a binary event occurring and predict the likelihood of all kinds of “yes” or “no” outcomes.**
* It assigns the probabilities to discrete outcomes using the Sigmoid function, which converts the numerical results into an expression of probability between 0 and 1.
* **The sigmoid function forms an S-shaped graph which means as x approaches infinity, the probability becomes 1 and as x approaches negative infinity, the probability becomes 0.**
* The model sets a threshold that decides what range of probability is mapped to which binary variable.
* **We can then decide the cut-off value at which we can differentiate between the different classes.**



* Types of Logistic Regression

1. **Binary** where the categorical response has only 2 outcomes (Yes or No).
2. **Multinomial** where the response can be three or more categories without ordering (Vegan, Non-veg, Vegetarian).
3. **Ordinal** where the response can be three or more categories with ordering (Movie Rating from 1 to 5)

**Advantages**

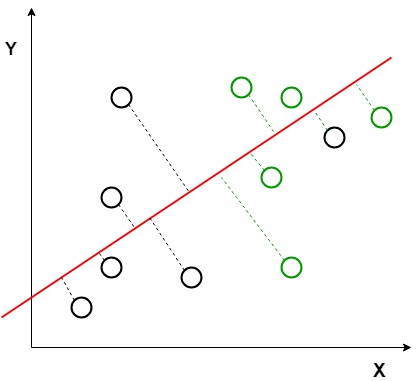
* It makes no assumptions about distributions of classes in feature space.
* The algorithm can easily be extended to multiple classes rather than the normal binary classes.
* Logistic regression is less inclined to over-fitting but it can overfit in high dimensional datasets.
* The predicted features and weights give inference about the importance of each feature and how the features are correlated.
* Logistic regression proves to be very efficient when the dataset has features that are linearly separable.
* It is used sometimes as a benchmark model to measure performance due to it relatively quicker and easier implementation process.
* The training time also comes out as far less than most complex algorithms.

**Disadvantages**

* If the number of observations is lesser than the number of features, logistic regression should not be used, otherwise, it may lead to overfitting.
* The major limitation of logistic regression is the assumption of linearity between the dependent variable and the independent variables.
* It requires average or no multicollinearity between independent variables.
* Data preparation can be tedious in logistic regression as both scaling and normalization are important requirements of logistic regression.
* It inherently runs on a linear model. This means even more restriction when it comes to implementing logistic regression as it won’t be applicable for nor-linear problems.
* Logistic regression isn’t immune to missing data unlike other machine learning models such as decision trees.

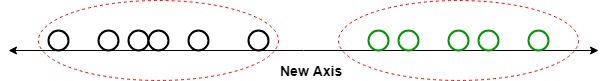
**Discriminant Analysis**

* It is a **statistical technique** used to **classify observation** into **non-overlapping groups** based on one more variable.
* It is also a **dimensionality reduction technique** as it reduces the features in higher **dimension space** into a lower dimension space.
* It tries to create a new axis in a way to maximize the separation between classes and project the data on it and by doing that, it reduces from using N features to N-1 features.
* Two criteria are used to create a new axis:
  1. **Maximize the distance** between means of two classes.
  2. **Minimize the variation** withing each class.
* Discriminant analysis uses a **linear discriminant function** that passes through the **centroids of the two groups** that can be used to discriminate between the two groups.
* Usually, the Discriminant analysis technique needs the **number of groups** minus one function to be able to split them to the number of groups.
* Each function includes **one or more attributes** that maximize the difference between the two groups.
* Each function alone split two groups to Default and Non-Default.



||

V



Advantages

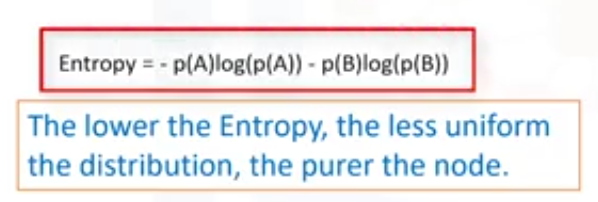
* It is simple, fast and portable algorithm. It still beats some algorithms (logistic regression) when its assumptions are met.
* Discriminant analysis minimizes the variance in the dataset by reducing the number of features.
* It is useful as it reduces the curse of dimensionality by effectively reducing the high-dimensional data into low-dimensional feature space.

**Disadvantages**

* It requires lots of unrealistic assumptions about the data.
* Does not give good results in case of unbalanced dataset.
* Not suitable for non-linear problems

**Decision Trees**

* DTs try to mimic human thinking steps by asking yes/no questions.
* The tree starts with a group of records and loops over this group trying to find an attribute and corresponding cut-off value that can be used to split the root node to two nodes.
* The split of nodes is based on measurements and values such as Entropy and Information Gain.
* Entropy is the measure of randomness and uncertainty which indicates the purity or impurity of the sets.



* Information gain is a measure of how much information a feature provides about a class and how uncertainty in the target variable is reduced, given a set of independent variables.



* The splitting of nodes tries to gather the similar records in the same node so, it tries to gather the good records in a node and the bad ones in the other.
* The purity is how much good vs bad in a node so, if 50-50, this won’t be a pure node.
* Each leaf should only contain mostly good or mostly bad data.
* One of the cons of using Decision trees is the high cost since each node requires field sorting. In other algorithms, a mixture of several fields is used at the same time, resulting in even higher expenses.
* Because slight changes in the data can result in an entirely different tree being constructed, decision trees can be unstable.

**Advantages**

* Easy to understand and interpret, perfect for visual representation.
* It doesn’t require a lot of preprocessing as there is no need for normalization, one-hot encoding and so on.
* Decision trees can automatically handle missing values and won’t lead to problems unlike other algorithms.
* They are usually robust to outliers and can handle them automatically.
* No assumptions are required about the distribution of the data.
* Feature selection happens automatically. So, unimportant features will not influence the result.
* Decision trees can be used to predict both continuous and discrete values so, they can be used for both classification and regression problems.
* Unlike logistic regression and other algorithms, they can capture nonlinear relationships so, they can be used to classify non-linearly separable data.

**Disadvantages**

* Decision trees tend to overfit. Due to the overfitting, there are very high chances of high variance in the output which leads to many errors in the final estimation and shows high inaccuracy in the results.
* It is very sensitive. Small changes in the data can affect prediction greatly. Adding a new data point can lead to re-generation of the overall tree and all modes need to be recalculated and recreated.
* Not suitable for large datasets as they can grow complex and lead to overfitting.
* The time complexity needed for operating this operation is very high and keep on increasing as the number of records gets increased. Decision trees with too many numerical variables take a lot of time for training.

**Credit Scoring and Default Risk Prediction: A Comparative Study between Discriminant Analysis & Logistic Regression**

Article in International Journal of Economics and Finance · March 2016

**Abstract**

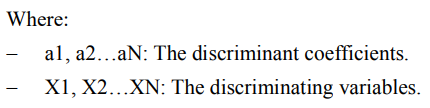
* This paper aims to develop models for predicting default risk of SMEs for one Tunisian commercial bank using Logistic Regression and Discriminant Analysis.
* They used a database containing 195 credit files granted to Tunisian SMEs which are divided into 5 sectors “industry, agriculture, tourism, trade and services” for a period from 2012 to 2014.
* The results proved that the two scoring techniques have a statistically significant power in predicting default risk.

**Features**

**Discriminant Analysis**

* We will be trying classify the companies to good or bad companies based on the features.
* To obtain an estimate for the model, they used a software called “SPSS.18”
* Healthy companies are rated by “1”
* Failing companies are rated by “0”

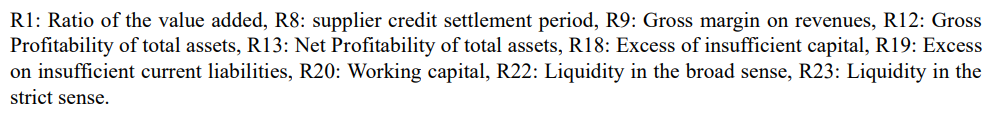




Using Wilks’s Lambda Test to prove the existence or non-existence of relationship between the dependent variable and the independent variables.

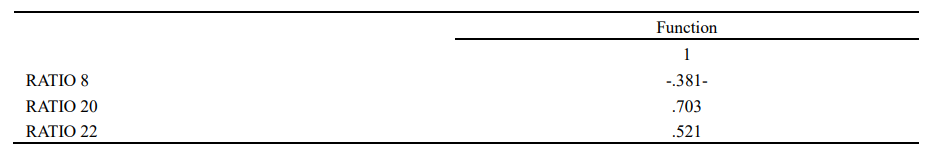
The tests value is always between 0 and 1 where 1 means the classes are equal while 0 means it has a great significance and could be used as the discriminating variable.

The model chose the following variables as the discriminating variables

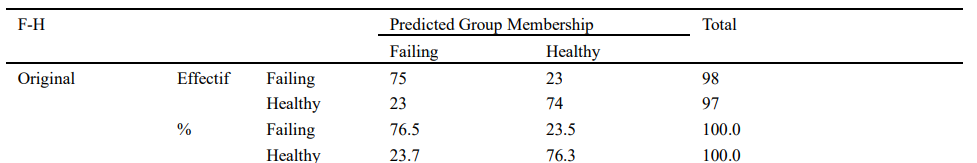


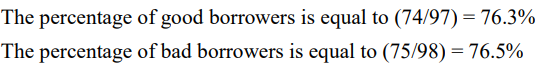
The model then chose the Discrimination Function and chose the weights for each of the chosen variables as the system tried to find from the chosen variables which ones have the greatest impact on the dependent variable and which ones maximizes the distance between the groups while minimizing the variance inside each group.

The model chose those 3 variables with their weights respectively

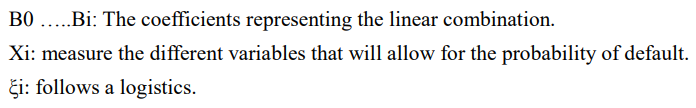


So, this is the score function which reflects the probability of the observation to fall in the default group



Logistic regression



Same conditions and target like we discussed in Discriminant analysis

