1. Literature Review on Credit Scoring

1.1 Literature background

1.1.1 History of Credit Scoring

Credit has been present in human history since the beginning of the civilisation, it can be dated back to 5000 years ago (Thomas et al., 2017). During the ancient Mediterranean period, the economy was fully based on agriculture, and it was the main source of wealth. Agricultural economies were characterised by having many ups and down due to seasonal changes, and this affected buyers' timing of payments. Some people had surplus of crops, while others had the need for this surplus but not the means to pay for it. Earliest references to banking and credit arrangements are exposed on stone tables in 2000 B.C (Lewis, 1992) as well as in the Code of Hammurabi around 1726 B.C, and it is fully extended until the end of the roman empire in which money was created, and credit started to be known as today (MacDonald & Gastmann, 2017).

Creditworthiness scoring has been among the oldest methods both in data analytics and risk management, and it is essential towards evaluating credit applications (Thomas et al., 2017). Credit scoring as we know it today dates to the 1950s. Back then there was no such thing as credit score. For people to get loans they had to hold long interviews with credit officers at the banks in which a decision was made based on individual judgement. Regarding this old system there were 2 big problems, first, the loan decision depended on a subjective judgement which was not an accurate way of determining whether a borrower will pay back a loan or not, second, the discriminatory bias regarding racial and gender characteristics of the applicants (Abdou & Pointon, 2011).

Although credit scoring as of today is quite new, the literature points out the existence of credit scoring attempts in the USA in the late 19th century during the creation of the first credit bureau in the east coast. They collected information about people to sell to landlords, retailers, anyone who can be interested. Most of the information collected was related to consumption habits, debts, gambling and drinking issues of the person in question (Lauer, 2017). However, the information was not statistically speaking analysed during the decision-making process, but more in a subjective way. Later, through the development of technology and the ability of companies and banks of gathering data some first quantitative approaches arose to credit scoring in the 1930s. Some retailers, banks and financial institutions introduced a point-based system for identifying who would and would not pay back a credit based on an application form that included information on the applicant’s occupation, age, race, marital status, income, neighbourhood of residence, etc. (Lauer, 2017). Also, the first records of statistical analysis on credit data were done by David Durant which used the discriminant analysis techniques to analyse instalment loan data and determine whether some loans where good or bad (Anderson, 2007).

In the 1950s Bill Fair and Earl Isaac created the Fair, Isaac and Company (FICO) with the aim of creating a impartial and standardised scoring system. During the launch of the credit scoring system only the American Investment Company decided to use the scoring method that was based on statistical analysis. There was however a huge resistance to use such system by banks since there were not the means to apply it out of the papers. However, it happened after the 1960s when companies started to computerise data from the customers and the FICO scoring system widespread throughout the USA in which some standardisations of defaults were brought into action, for example, the ones to mark late payments as 30, 60 or 90 days behind. By this time most of credits were granted to companies and the share of granted credits to individuals was too small.

During the 1960s and the arrival of credit cards, banks realised the importance of credit scoring, the number of new applicants for credit cards made it impossible for banks to have the credit scoring as a non-automatised lending decision. The growth of computing power made popular the credit scoring adoption by the 1980s and used the system not only for credit cards but for other financial products for example, mortgages, personal loans, etc. Also, by this time new statistical techniques were introduced, more specifically, the logistic regression and linear programming. Those techniques are still being used, however, as computers and technology develops, newer and more precise techniques are put into action nowadays that provide higher quality results in terms of credit scoring, for example, artificial intelligence (AI) and machine learning (ML).

1.1.2 What is Credit Scoring

Credit scoring can be defined as a set of statistical techniques in a form of a set of decision models that aims to determine whether a lender will grant a loan to a borrower. These techniques are used to establish how much of the loan credit a borrower should get and magnify the profitability of the lenders (Thomas et al., 2017). Credit scoring helps to assess the level of creditworthiness of the borrower by determining the probability of default to a certain loan.

Being creditworthy is not an attribute of a certain person, however, it is an assessment done by the lender when evaluating the profile of the borrower by the usage of credit scoring techniques. In other words, credit scoring determines the probability that a borrower will be “good” or “bad” (creditworthy or uncreditworthy) based on the profile of the borrower, as well as the economic scenario, potential loses, churns, and approval rates. All these factors together are relevant in the credit adjudication process (Siddiqi, 2013).

Businesses’ mission is to create value and maximise profits (Handy, 2002). Identify those “good customers” is very important in the financial industry, to this, banks and risk managers constantly use credit scoring techniques to select those “right” customers (low-risk customers) and implement a marketing strategy to offer financial products and get some profit. It is important to remark that the main source of revenue of banks is lending money and charging an interest rate to the borrower based on the probability of default of the loan (risk), this interest rate in known as the price of money, therefore the higher the risk of the loan the higher the interest rate related to it (Lee & Hogarth, 2018). For banks to maximise profits they need to lend as much as possible to low-risk customers, or to reduce cost related to the lending process by limiting granting loans to customers with a high probability of default (high-risk). The problem is that sometimes these low-risk customers do not need loans or financial products and they reject the banks offers regardless the effort put in the marketing campaign (Siddiqi, 2017), and that’s why banks and financial institutions make a huge effort in identifying those high-risk customers that may default a loan.

Banks and financial institutions have a huge history of data of their customers, along with the application forms when applying for a loan. The information gathered by the banks is presented generally as a form of a scorecard where the characteristics of the profile of a given customer (borrower) have a score, and the sum of these scores determines whether a loan will be granted or not to a customer. Generating a scorecard is very important process for a bank and for this many statistical and data mining techniques are used (Thomas et al., 2017). Ultimately, credit scoring is used to facilitate decision-making in business, although it is often associated with the statistical techniques and processes used in the scorecard's development.

1.1.3 What are Scorecards

As explained before, credit scoring uses predictive statistical methods to classify borrowers by their probability of being a "good" or "bad" in the future, based on the lender's past experiences and the profile of the borrower. Credit scoring models are presented in different forms, but the most are presented in form of a regression (Anderson, 2007).

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n + e \qquad (1)$$

Where $\boldsymbol{y}$ is the dependent variable associated with the outcome that refers to the probability of default, being 0 for "bad" and 1 for "good" customer, it may also come as a form of "logit" or "probit" depending on the model specifications. Being $\boldsymbol{x}$ is the set of independent variables as a form of an original, transformed, or dummy variable. $\boldsymbol{b}$ stands for the regression coefficient by which the independent variables are weighted indicating the relative importance and $\boldsymbol{e}$ is the regression error which cannot be captured by the model.

Regression coefficients are used to present a model that attempts to explain the relationship between the independent variables $\boldsymbol{x}$ and the independent variable $\boldsymbol{y}$**.** Traditional scorecards use classed variables in which scores are given if certain condition holds true based on the model coefficients, for example:

Condition	Action	Score
Age < 25	Deduct	20
Age > 40	Add	25

If owning a house	Add	40
If renting a house	Deduct	15
If married	Add	20
If single	Deduct	20
etc	…	…

Source: own work

The scorecards can be also presented in a tabular format as shown in the next table in which the characteristics are compressed, and attributes are columns assessed on score (points) inspired on a FICO scorecard:

Characteristic	Attributes					Points
Years at address	<3 years 30	3 – 6 years 36	>6 years 38		Blank 20	30
Years at employer	<2 years 30	2-8 years 39	9-20 years 43	>20 years 69	Blank	43
Accommodation status	Own 41	Rent 32	Parents 35		Other 36	32
Civil status	Married 40	Single 22	Divorced 32		Blank 32	32
Past experience	None 3	New 15	Updated 30	Past due -5	Write-off Reject	15
					Final Score	152

Source: based on Anderson, 2007.

The scorecard characteristics are usually obtained from different sources of data to which the lender has access during the application process. For building a scorecard, it is considered behavioural, financial, socio-economic, and demographic data (Martin & Evzen, 2006). Regarding the sources of this data, most of the data used for evaluating the application comes internally, from the application form, credit bureaus, public commercial registers, financial statements, tax statements, integrators of economic information, and national statistical offices (Kaszyński et al., 2021). Naturally, the variables to consider may vary if the loan applicant is an individual or a company. To this, Kaszyński et al. (2021) defines a list of potential variables to include in the traditional scorecard depending on the type of applicant. For individuals, it is important to know the main source of income, disposable income, alternative sources of income, type of activity, stability of fixed expenses, assets, saving rate, credit history, delinquencies, and geolocation. However, for companies a lender would be also interested in knowing the nature of business, business history, management board, balance sheet, income statement, cashflow, suppliers, clients, delinquencies, and exposure to cross border risks.

Based on the score that a borrower gets from the application form a lender may decide to grant a loan or reject it. In particular, the total score is a measure of the risk associated with the loan for an application, and the lender (depending on if is risk adverse or risk lover) may grant high-risk loans that comply with a certain score cut off to which higher interest rate will be charged, or simply reject it.

For example, based on the previous table, the lender may decide to reject all loans that score less than 150 points, those between 150 – 160 points may be charged a higher interest rate, and those will a score higher than 160 will be granted a loan automatically. For those high-risk applicants, the lender doesn't necessarily need to charge a higher interest rate, but may request extra conditions such as assigning a lower loan/credit, charge a higher premium on insurance, ask for a default insurance, request for extra documentation on assets, etc. The "due diligence" policy may depend on the expected approval rate and revenue or profit potential at each risk level (Siddiqi, 2017).

The attributes in the scorecard are assigned points based on statistical techniques. There are different ways to calculate the weights of the attributes in a scorecard depending on the kind of credit scoring model used, and these models provide the predictive strength of the characteristics. To this, the literature defines 2 kind of predictive modelling techniques in credit scoring: parametric modelling that makes assumptions on the data, and non-parametric modelling which doesn't make any assumptions on the data. (Abdou & Pointon, 2011; Anderson, 2007; Lauer, 2017). Also, there are other factors considered when calculating the scores of the attribute, such as the correlation between them, and operational factors (Siddiqi, 2013). Lastly, the total score is calculated by summing all scores of the attributes of the applicant.

1.1.4 Credit scoring modelling characteristics

Different statistical methods have arisen over the years, at the very beginning it was mentioned that basic statistical tools were used for credit scoring, and more sophisticated statistical methods were taking over. Actually, most of the techniques used in credit scoring techniques are within the field of predictive modelling and these can be classified in 2 sections:

- Parametric modelling techniques: The main feature about this type of modelling techniques is that they make assumptions about the data such as Linear Discriminant Analysis and Logistic Regression. The main benefit of these models is that they are simple and easy to interpret and understand, they are also fast computationally speaking, and require less data for training purposes. Although these assumptions can help us to get more interpretable estimates, they also limit what can be learned. These models are highly constrained, have a limited complexity, and it is important to mention that the literature points out that these methods many times fail to match very complex underlying processes (Puertas et al., 2003)
- Non-parametric modelling techniques: These models don't make assumptions on the underlying process and have more flexibility in terms of adapting any functional form from the training data. Some these models are Pondered Regression, Regression Trees, Algorithm C4.5, Multivariate Adaptative Regression Splines, Random Forest, Extreme Gradient Boosting, Support Vector Machines, and Neural Artificial Nets (Kennedy, 2013; Knutson, 2020; Puertas et al., 2003).

1.2 Best practices in Credit Scoring

Credit scoring heavily depends on data and during the scorecard development there are many data considerations. With the collected data the lender can build a credit scoring model that aims to predict future default events by using statistical methods. When building a credit scoring model some factors must be taken into account so the reliability of the final model will not be compromised. In this sense, the literature highlights 4 important factors to consider (Abdou & Pointon, 2011; Anderson, 2007; Martin & Evzen, 2006; Onay & Öztürk, 2018; Siddiqi, 2017):

- Transparency: The information can be used for assessment,
- Structure: The data is easy to analyse data considering its form,
- Data quantity: There are enough observations for model development,

- Data quality: consistency and accuracy of the data.

Anderson (2007) defines the concept of best practices as "…processes, techniques, methodologies, and the use of technology, equipment, and resources that have proven record of success at providing a desired result". In order to ensure the quality of the credit scoring model the data must be pre-processed and analysed in advanced so that we can make sure of the precision of the estimations. Some practices and factors are industry specific. In credit scoring modelling, some general best practises in credit scoring modelling are explained by Kaszyński et al., (2021) as follows:

Firstly, for building a reliable credit scoring model we need very large datasets to ensure that our population sample for training a model can converge to the real population and therefore we can get more accurate estimates (Martin & Evzen, 2006). For this it is important to understand the target population, meaning that when building a scoring model, the scorecards shouldn't be separated based on different products (credit card, cash loans, etc) or segmentate the scorecards based on the characteristics of the type of applicant (size of company, balance sheet, etc). This can lead to a more complex architecture, and it lowers the number of observations in each of the specific models, and therefore it affects the model stability. However, the target population should be defined considering the available information avoiding the loose of observations due to segmentation. The idea is to treat certain atypical observations (clients) in the training sample of the model as "outcome exclusions" and minimise the amount of observations that may distort the model.

Secondly, we must understand the data and assess the kind of information we get for training a scoring model. The financial and banking sector in general is vastly regulated and certain rules apply for processing and modelling the data (Knutson, 2020; World Bank Group, 2019). It is important that the training model has as many relevant explanatory features that helps to correctly assess the probability of default of a customer. Although the intuition is to include as many variables as possible, we must pay extra care when handling the data and avoid violating the privacy of the customer. European authorities warrant customer's right to explanation and acknowledge of data processing in the latest data protection regulation (*EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679*, 2016). To this, customers must be informed that their data is being handled and assessed. Although this principle applies in the European Union, some countries are applying a similar legislation, therefore we need to make sure that our training sample does not contain data which does not comply to the current legislation (Demajo et al., 2020). If some “sensitive” information is detected in the training sample that does not comply to the legislation, then it should be removed.

Thirdly, we must keep in mind the data feed process and understand the how the data is updated in the database. The training data generally contains a combination of internal data as well as data from external sources (as explained above) and this external data in some cases can be not updated. Therefore, the training dataset must be prepared in a way that these time differences are spotted and reflected. As a rule of thumb, credit scoring models must be trained with the same data feed in which they will be used in production. Furthermore, is important to keep track of the data feed an ensure the inflow of up-to-date data and avoid situations when credit bureaus stop data feed to subscribers that breach the reciprocity agreement (Anderson, 2007).

Fourthly, we must remove visible rubbish from the data set in the sense that it must be included for training purpose those variables that are strictly relevant for modelling issues and regulatory standards. A few examples of variables that are irrelevant are those with variable coverage in time, optional variables that may come from the application form, dummy variables that indicate missing data, categorical variables with many non-hierarchical categories, customer identification data, unprocessed transaction data, and data compromising the privacy of the customer. A good practice is to create summary variables for each unprocessed transaction data that contains count and frequencies.

Fifthly, there must be a tight control on the quality of the outcome variable. Some of the good practices addressed by Kaszyński et al., (2021) in this issue are the following: get a clear definition of delinquency, mind that a restructured loan is usually a bad loan, First-in-first out (FIFO) is a better definition for a bad days past due than Last-in-first out (LIFO), there are alternative indicators for bad behaviour according to the European Bank Authority guidelines for products with compromised days past due, better to use a less precise “bad” definition than a more precise one but less consistent, the outcome variable must be defined taking into account all products on the customer level, and lastly, inactive clients should be marked as outcome exclusions.

Finally, it is important to remark the importance of Basel II compliance, and International Financial Reporting Standards (IFRS) 9. This issue will be further discussed in the next chapters.

1.3 Classical credit scoring approach

1.4 Drawbacks and obstacles of classical scoring approach

1.5 Economics behind credit scoring

2. Machine learning and AI in credit scoring

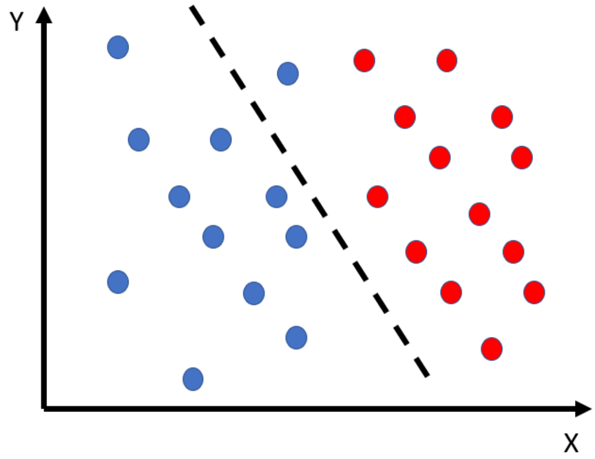
As explained in previous sections, machine learning techniques are widely used in credit scoring due to the simplification of complex scoring decisions, aiming to generate real-time predictions and reduce credit risks. In this chapter, we will dig further into the most common modelling techniques used nowadays in credit scoring, as well as the main problems that may arise in this process.

2.1. Parametric models used in credit scoring

Nowadays credit-scoring systems are based on statistical and operational research methodologies that allow automatization of the scoring process increasing productivity and predicting defaults in a faster way. They are among the most effective and lucrative uses of statistical theory (Langdon et al., 1992). In previous sections it was mentioned 2 different approaches used in credit scoring modelling. While parametric modelling techniques make several assumptions about the underlying data, non-parametric techniques make few (Anderson, 2007). The parametric models, including linear discriminant analysis, logistic regression, and regularisation methods (such as ridge regression, lasso regression and elastic net), are discussed below.

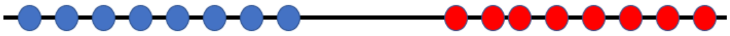
2.1.1. Linear discriminant analysis

Linear Discriminant Analysis (LDA) is a widely used technique for dimensionality reduction and classification. LDA enables class separability by establishing a decision region between the various classes by maximising the distance between the means of the classes, and by minimising the between-class variance and the within-class variance (Mohanty et al., 2013). Sir Fisher R. A. (1936) developed the linear discriminant function, trying to identify the set of factors that best divided two groups using accessible attributes, and LDA came later as a simple generalisation of this function. In credit scoring, the two categories are those categorized as good and bad customers by the lender and the characteristics are the details in the application form and credit bureau information.



Source: own work

The above graph is an example of credit scoring in which the classes (good and bad customers) are separated by the dashed line defined by LDA. These data points are projected into the dashed line reducing this two-dimensional graph into a one-dimensional graph. That's why this method can be considered as dimensionality reduction algorithm. However, when the mean of the distributions is the same (or shared) this algorithm fails to convert the n-dimensional data into a single one-dimensional graph that makes both classes separable such as:



Source: own work

From a mathematical point of view, LDA projection can be applied to a function if

$$Y = w_0 + w_1 X_1 + w_2 X_2 + \ldots + w_p X_p \tag{2}$$

This function represents a linear combination of observations, then the total sum of squares of the dependent variable Y can be calculated as follows:

$$Y^T HY = a^T X^T HXa = a^T Ta \tag{3}$$

With a centering matrix around $H = X^T HX$

Suppose we have samples $X_j,\ j = 1, \ldots, J$ from J populations. Fisher's suggestion was to find the linear combination that maximizes the ratio of the between-group-sum of squares to the within-group-sum of squares. The within-group-sum of squares measures the sum of variations within each group, whereas the between-group-sum of squares measures the variation of the means across groups. The within-group-sum of squares is defined as follows:

$$\sum_{j=1}^{J} \mathcal{Y}_j^\top \mathcal{H}_j \mathcal{Y}_j = \sum_{j=1}^{J} a^\top \mathcal{X}_j^\top \mathcal{H}_j \mathcal{X}_j a = a^\top \mathcal{W} a \tag{4}$$

And the between-group-sum of squares is defined as follows:

$$\sum_{j=1}^{J} n_j \left(\bar{y}_j - \bar{y}\right)^2 = \sum_{j=1}^{J} n_j \left\{a^\top \left(\bar{x}_j - \bar{x}\right)\right\}^2 = a^\top \mathcal{B} a \tag{5}$$

The total sum of squares from equation (3) is calculated as the sum of the within-group-sum of squares and the between-group-sum of squares:

$$a^\top \mathcal{T} a = a^\top \mathcal{W} a + a^\top \mathcal{B} a \tag{6}$$

The objective is to maximise (optimise) the ratio of the between-group-sum of squares to the within-group-sum of squares as mentioned before.

$$\frac{a^\top \mathcal{B} a}{a^\top \mathcal{W} a} \tag{7}$$

The idea is to get the eigenvector of $\mathcal{W}^{-1}\mathcal{B}$ that maximises the previous ratio that corresponds to the largest eigenvalue. In credit scoring, we have 2 classes in our dependent variable associated with default and no-default. With $n1$ and $n2$ observations per class.

$$\mathcal{B} = \left(\frac{n_1 n_2}{n}\right) d d^{\top} \tag{8}$$

Where $d = (\bar{x}_1 - \bar{x}_2)$. Then we obtain the only eigenvalue from $\mathcal{W}^{-1}\mathcal{B}$ and the eigenvector which is $a = \mathcal{W}^{-1}\mathcal{B}$. Once we calculate the eigenvector, we can set decision rule as follows:

$$\begin{aligned} x \to \Pi_1 \quad \text{if} \quad a^{\top}\left\{x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)\right\} > 0 \\ x \to \Pi_2 \quad \text{if} \quad a^{\top}\left\{x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)\right\} \leq 0 \end{aligned} \tag{9}$$

Based on that, we can highlight that linear discriminant analysis (LDA) uses the training observations to determine the location of a boundary between the response classes. The boundary location is determined by treating the observations of each class as samples. Theoretically, we could fit an (n) dimensional normal distribution from a multidimensional normal distribution to the observations in each class. Calculating this involves calculating the mean vector and covariance matrix for each class, as these determine the centre and shape of the distribution, respectively (Johnson & Wichern, 2002; Thomas et al., 2017).

Having fitted the distributions, we could draw a boundary between the classes by selecting the set of points where the probabilities are equal. Observations on one side of the boundary would be classified as one class, and observations on the other side as the other class. We can do all this theoretically, resulting in an equation for the boundary that depends on the parameters of the fitted distribution. This means that we don't need to go through the whole process to perform discriminant analysis; we need to calculate the means and covariances and apply the formula for the boundary.

2.1.2. Logistic regression

Logistic regression is by far the most widely used algorithm in credit scoring, its usage is preferred due to the simplicity and ease of interpretation. Therefore, a huge section of this dissertation is dedicated to cover the logistic regression model.

2.1.2.1 Origins of logistic regression

The logistic function was developed in the nineteenth century to describe population growth and the behaviour of autocatalytic chemical processes. Now, the growth rate of a quantity $W(t)$ over time is given by:

$$\dot{W}(t) = \frac{dW(t)}{dt} \tag{10}$$

Assuming that $\dot{W}(t)$ is proportional to $W(t)$ we get:

$$\dot{W}(t) = \beta\, W(t), \beta = \dot{W}(t) / W(t) \tag{11}$$

where β refers to the constant rate of growth, thus leading to the exponential growth model

$$W(t) = A\, e^{\beta t} \tag{12}$$

where the initial value $W(0)$ may replace A.

However, the exponential growth model is devoid of any upper limit. This problem was approached by the Belgian astronomer and statistician Alphonse Quetelet (1795-1874) and his student Pierre-Francois Verhulst (1804-1849) through the inclusion of an additional term, representing the increasing resistance to further growth, in equation (2) as follows:

$$\dot{W}(t) = \beta\, W(t) - \phi\, (W(t)) \tag{13}$$

Experimentation with varied forms of ϕ led to the following model when ϕ is a quadratic function:

$$\dot{W}(t) = \beta\, W(t)\, (\Omega - (W(t)) \tag{14}$$

Where Ω refers to the upper limit or saturation level of W.

In the above model, $\dot{W}(t)$ is proportional to both $W(t)$ and $(\Omega - (W(t))$. Expressing $W(t)$ as a proportion $\{P(t)\}$ of Ω, or $P(t) = \frac{W(t)}{\Omega}$, which results in the following differential equation:

$$P(t) = \beta\, P(t)\, \{1 - P(t)\} \tag{15}$$

Solving the above differential equation, we get the final logistic function:

$$P(t) = \frac{e^{\alpha + \beta t}}{1 + e^{\alpha + \beta t}} = \frac{1}{1 + e^{-(\alpha + \beta t)}} \qquad (16)$$

Equation (7) was named as the ***logistic*** function by Verhulst. In regression analysis, α and β may be interpreted as the intercept and the regression coefficient (or the slope of the regression line), respectively.

The logistic function was rediscovered in 1920 by Raymond Pearl (1879-1940), the Director of the Department of Biometry and Vital Statistics at John Hopkins University, and his deputy Lowell J. Reed (1886-1966) while studying the United States' population increase (Cramer, 2002).

2.1.2.2. The logic of logistic regression

Discrete or qualitative rather than continuous or quantitative events are common in many social phenomena; for example, an event happens or does not happen, a person's life can be changed in a variety of ways that involve a characteristic, event, or choice, or large social entities such as groups, organizations and nations can arise or disintegrate, become insolvent, confront, revolt, and so on. A dichotomous indicator or dummy variable is the most common way to represent discrete binary phenomena (Pampel, 2000).

On the surface, a binary qualitative dependent variable seems to be appropriate for use in multiple linear regression. The dependent variable assumes only two values of zero and one. However, the estimated values for regression are in the form of mean proportions or probabilities conditional on the values of the independent variables. The regression coefficients may be interpreted as the increase or reduction in the estimated probability of possessing a characteristic or experiencing an event due to a unit change in the independent variables. However, such linear regression faces the following problems:

1. **Probability ceilings:** Probabilities and proportions, by definition, cannot exceed one (ceiling) or fall below zero (floor). Nonetheless, the linear regression line may stretch upward into positive infinity (or extend downward towards negative infinity) as the values of the independent variables increase (or decrease). A model may provide illogical and useless predicted values of the dependent variable greater than one and

less than zero based on the slope of the regression line and the observed values of the independent variables.

2. **Additivity assumption:** Typically, linear regression assumes additivity, which states that the influence of one independent variable on the dependent variable remains constant regardless of the values of the other independent variables. Although models may include chosen product terms to accommodate non-additivity, a dummy dependent variable will almost certainly break the additivity assumption for all possible combinations of the independent variables. When the value of one independent variable reaches a level sufficient to push the probability of the dependent variable close to one (or close to zero), the impacts of other variables have little effect. Thus, the ceiling and floor impose an intrinsic, non-additive and interactive nature on the impact of all independent variables (Pampel, 2000).

3. **Normality and homoscedasticity assumption:** Linear regression with dichotomous dependent variable violates the normality assumption (each value of the independent variables in the population is associated with a normal distribution of error terms around the predicted value of the dependent variable) and the homoscedasticity assumption (the dispersion of the error terms for each value of the independent variables is similar) (Pampel, 2000).

The relationship between the probability of a dummy dependent variable and an independent variable is inherently non-linear. A constantly changing curve, such as the S-shaped curve, represents the relationship more smoothly and adequately than a straight line. Although other non-linear functions may depict the S-shaped curve, the logistic or logit transformation has gained popularity due to its desirable characteristics and relative simplicity.

Let Y_i be the observations (i=1,2,3,...,n) of a dichotomous dependent variable (Y) representing some event and assuming the values of one (occurrence of the event) and zero (non-occurrence of the event) only. Given the probability of the occurrence of the event as $P\ (Y_i = 1) = P_{\mathrm{i}}$, the probability of the non-occurrence of the event will be $P\ (Y_i = 0) = (1 - P_i\ )$. In his paper, Pampel (2000) explains that the odds of the probability of the occurrence of the event relative to the probability of the non-occurrence of the event is then given by:

$$O_i = \frac{Pi}{(1 - Pi)} \tag{17}$$

Based on the above formula we can derive the following:

If $Pi = 0$ then $O_i = 0$

If $0 < P_i < 1$ then $0 < O_i < \infty$

If $P_i = 1$ then $O_i = \infty$ therefore $0 \leq O_i \leq \infty$.

Odds are generally expressed implicitly as a ratio to one or as a single number. For instance, if the probability of an event equals 0.4, the odds are (0.4 / 0.6) or 0.667, indicating that the event occurs 0.667 times for each time it does not occur, or 667 occurrences per 1000 non-occurrences. Even though both probabilities and odds have a lower bound (floor) of zero, denoting the increasing likelihood of an event with increasingly large positive numbers, odds have no upper bound (ceiling), unlike probabilities. The transformation of probabilities into odds eliminates the probabilities' ceiling value of one.

Now, the logit or logged odds is formed by taking the natural logarithm of the odds to eliminate the odds and hence probabilities' floor value of zero, as:

$$L_i = ln\, O_i = ln\, [\frac{Pi}{(1 - Pi)}] \tag{18}$$

Based on the above formula we can derive the following:

If $O_i = 0$ then $L_i = \infty$

If $0 < O_i < 1$ then $L_i < 0$

If $O_i = 1$ then $L_i = 0$

If $1 < O_i \leq \infty$ then $0 < L_i \leq \infty$ therefore $-\infty \leq L_i \leq +\infty$

Taking the above formula in consideration, the non-linear relationship between the independent variable (X) and the probability of the occurrence of the dichotomous dependent variable (Y) conditional on the observed values of X (X_i) expressed as P (X_i) = [P (Y_i = 1 | X = X_i)], may be transformed into a linear regression function based on the logit of P (X_i) from equation (2.2) as:

$$L\,[\,P\,(X_i)\,] = ln\,[\frac{P(Xi)}{(1-P(Xi))}] = \alpha + \beta X_i \qquad (19)$$

where α and β denote the intercept and the regression coefficient (or the slope of the linear regression line), respectively. Equating equations (1.7) and (2.3) and replacing t with $X_{i,}$ we get:

$$P(X_i) = \frac{e^{\alpha+\beta X_i}}{1+e^{\alpha+\beta X_i}} = \frac{1}{1+e^{-(\alpha+\beta X_i)}} \qquad (20)$$

2.1.2.3. Interpretation of logistic regression coefficients

The sign of β in equations (2.3) and (2.4) determines whether P (X) is increasing or decreasing as X increases. The rate of ascent or descent increases as the magnitude of β, that is $|\,\beta\,|$, increases; as β approaches zero, the curve flattens to a horizontal straight line. When β equals zero, Y is independent of X. For quantitative X with β greater than zero, the curve for P (X) has the shape of the cumulative distribution function of the logistic distribution, given by:

$$F\,(X) = \frac{e^{\frac{(X-\mu)}{\tau}}}{1+\ e^{\frac{(X-\mu)}{\tau}}} \qquad (21)$$

Where μ and τ are the mean and standard deviation respectively of the logistic distribution. Due to the symmetrical distribution of the logistic density, P (X) approaches one at the same rate that it approaches zero (Agresti, 2002). Multiple interpretations exist for β in terms of logged odds, odds and probabilities and the nature of the independent variable being continuous or dummy (Pampel, 2000). We can focus on the following ways to interpret logistic regression coefficients:

1. Logged odds: In the case of continuous independent variables, the logistic regression coefficients indicate the change in the projected logged odds of the occurrence of an event for one unit change in the independent variables. In contrast, an implicit comparison of the indicator group with the reference or excluded group is made by a one-unit change in the case of dummy independent variables. Browne (1997, p. 246), for example, uses logistic regression to forecast labor force participation for 922 female heads of home between the ages of 18 and 54 in 1989. For the continuous independent variable ‘Years employed’, the logistic regression coefficient of 0.13 indicates an increase in the logged odds of the dependent variable ‘Labor Force Participation’ by 0.13 for an additional year of employment. The author also compares the ‘Labor Force Participation’ with two dummy variables, namely ‘High school dropout’ and ‘High school graduate’ to that of the reference group consisting of women with some college education. These two dummy variables have correlations of -1.29 and -0.68, respectively, indicating that the logged odds of being in the labor force are 1.29 lower for high school dropouts than for those with some college and 0.68 lower for high school graduates than for those with some college education (Pampel, 2000).

2. Odds: Let us consider the following binary logistic regression model with two independent variables, X_1 and X_2 :

$$L\,[\,P\,(X_i)\,] = ln\,[\frac{P(Xi)}{(1-P(Xi))}] = \alpha + \beta X_1 + \beta X_2 \tag{22}$$

Where α is the intercept term, and β_1 and β_2 are the regression coefficients. Exponentiating both sides we get:

$$e^{(P(Xi)/1-P(Xi))} = e^{\alpha+\beta_1 X_1+\beta_2 X_2} \tag{23}$$

$$\frac{P(Xi)}{(1-P(Xi))} = e^{\alpha} * e^{\beta_1 X_1} * e^{\beta_2 X_2}$$

The equation (3.2) determining the odds is multiplicative, where the predicted value of the dependent variable does not change when multiplied by a coefficient of one. Therefore, the effect of each independent variable on the odds can be measured by taking the antilog of the regression coefficients. The percentage change in the odds for a one-unit change in the independent variable (with regression coefficient ‘β‘) which is given by:

$$\%\Delta = (e^{\beta} - 1) * 100 \tag{24}$$

Now it presents a more meaningful interpretation of the regression coefficients. With reference to Browne's study mentioned above, the exponentiated coefficient for the continuous variable ‘Years employed’, $e^{0.13}$ or 1.14 indicates that a one-year increase in employment multiplies the odds of the dependent variable ‘Labour Force Participation’ by 1.14 or increases the odds by a factor of 1.14 or 14%. In the case of the dummy variable

‘High school dropout’, $e^{-1.29}$ or 0.28, indicates that a one-unit increase in the variable multiplies the odds of ‘Labour Force Participation’ by 0.28 or the odds are 0.28 times or 72% smaller than those with some college education (reference group) (Pampel, 2000).

3. Probabilities: Due to the non-linear and non-additive nature of the interactions between the independent variables and probabilities, they cannot be adequately described by a single coefficient. Instead, the influence on probability distributions must be determined for a given value or group of values, the choice of which is determined by the researcher’s concerns and the nature of the data (Pampel, 2000).

Now let’s break down on a more concise explanation for the interpretation of the coefficients estimates in logistic regression.

In the case of continuous variables, calculating the linear slope of the tangent of a non-linear curve at every single point is a simple approach to determine the effect of a continuous independent variable on probability. The slope of the tangent line is given by the partial derivative of the non-linear equation linking the independent variables to the probabilities. The partial derivative, which measures the change in the probability for an infinitesimally small change in X_k (k = 1,2) and defines the slope of the tangent line or the change in the tangent line due to a one-unit change in X_k at that value, is obtained from equation (3.1) as:

$$\frac{\partial P}{\partial Xk} = \beta_k * P * (1 - P) \tag{25}$$

Considering Browne’s study mentioned above, the logistic regression coefficient for ‘Years employed’ equals 0.13; the mean of the dependent variable, the expected probability of ‘Labor Force Participation’ equals 0.83; and the probability of not participating equals 0.17. From equation (3.4), the partial derivative is given by (0.13 * 0.83 * 0.17) or 0.18, which indicates that an increase of one year of employment increases the probability of participation by 0.018 or almost 2% at the mean. The marginal effect reaches its maximum value of 0.032 when P = 0.5.

Long (1997) discusses several alternative methods to present a more complete summary of the range of effects of an independent variable on probabilities (Pampel, 2000). However, disagreements exist among analysts on the usefulness of even calculating a single partial derivative given the constraints of describing a non-linear and non-additive relationship with a single coefficient (DeMaris et al., 1990; DeMaris, 1993; Roncek, 1993).

In the case of dummy independent variables, the tangent line for infinitesimally small changes makes little sense. However, it is possible to compute expected probabilities for each of the two groups and then measure the group differences in probabilities by subtracting the two probabilities. The steps involve the following calculations:

1. Logit for the omitted group: $L_o = ln\ (P_o\ /\ (1 - P_o))$ (26)
2. Logit for the dummy variable group: $L_d = L_o + \beta_d$ (27)
   where β_d is the logistic regression coefficient
3. Probability for the dummy variable group: $P_d\ = \{1\ /\ (1 + e^{-L_d})\}$ (28)
4. The difference in probabilities: $P_d - P_o$ (29)

With reference to Browne's study mentioned earlier, the above calculations using the mean or expected probability of the dependent variable 'Labor Force Participation' (P_o = 0.83), regression coefficient of 'High school dropout' (b_d = –1.29), and women with some college education as the omitted group, are shown below:

Logit for women with some college education: L_o = ln (0.83 / 0.17) = 1.586
Logit for 'High school dropout': L_d = 1.586 –1.29 = 0.296;
Probability for 'High school dropout': P_d = {1 / (1 + 0.7438)}= 0.573;
Difference in probabilities: $P_d - P_o$ = 0.573 – 0.83 = 0.257
Thus, high school dropouts have a probability of participating which is 0.257 lower than those with some college education (Pampel, 2000).

It is important to mention that predicted probabilities may be used as the partial derivative for continuous variables, just as they can be used for dummy variables. However, changes in projected probability reflect the real impact of a discrete change in the independent variable *X*, for example, one unit, rather than the influence on the tangent line suggested by an instantaneous or infinitesimally small change in *X* (Pampel, 2000). As a result, some favour using projected probabilities over the partial derivative when dealing with discrete changes (Kaufman, 1996; Long, 1997). The predicted probabilities are calculated using the same procedures for dummy variables, except that *X* is substituted for the omitted group and (*X* + 1) for the dummy variable group. The following sequential steps may be followed:

1. Calculating Logged odds of *P* (i.e., the logit before the change in X)
2. Adding *X*'s logistic regression coefficient to the starting logit and computing the probability for the new logit
3. Subtracting the starting probability (at *X)* from the second probability (at *X* + 1) shows the effect of a one-unit change in *X* on the predicted probability at *P*

Considering the above-mentioned Browne's study, we see *P* = 0.83, the coefficient for 'Years employed' = 0.13, the logit at *P* = ln (0.83 / 0.17) or 1.586. Adding the coefficient to this logit yields 1.716 (1.586 + 0.13). From equation (3.7), the probability for (Years employed + 1) is calculated as 0.848. The difference between 0.848 and 0.83 equals 0.018 which indicates that a one-year increase in 'Years employed' increases the probability of 'Labor Force Participation' by 0.018 at it's mean (Pampel, 2000).

2.1.2.4. Estimation and model fit of binary logistic regression

In the case of a binary logistic regression model, we can estimate the coefficients using different methods. The most popular in the field of credit scoring and machine learning are the following:

1. Maximum Likelihood Estimation (MLE): Due to the binary nature of the dependent variable, the error term does neither have a normal distribution nor equal variances for the values of the independent variables. As a result, the estimating approach derived from the Ordinary Least Squares (OLS) criterion, which involves minimizing the sum of the squared deviations between the observed and predicted values of the dependent variables, fails to produce efficient estimates. Therefore, rather than using OLS, logistic regression depends on maximum likelihood algorithms to estimate the coefficients.

For logistic regression, the estimation of the regression coefficients begins with an expression for the likelihood of observing the pattern of occurrences (*Y* = 1) and non-occurrences (*Y* = 0) of an event or characteristic in a given sample. This expression, termed the likelihood function, depends on unknown logistic regression parameters. Maximum Likelihood

Estimation (MLE) finds the model parameters that provide the maximum value for the likelihood function, thereby identifying the estimates for model parameters that are most likely to give rise to the pattern of observations in the sample data. The maximum likelihood function in logistic regression is given by:

$$LF = \prod \{ P_i^{Y_i} * (1 - P_i)^{1 - Y_i} \} \qquad (30)$$

Where *LF* refers to the likelihood function; Y_i refers to the observed value of the binary dependent variable for case i ; $P_{i,}$ which refers to the predicted probability for case i is given by $Pi = \{ 1 / (1 + e^{-L_i})$, and where L_i is the logged odds determined by the unknown regression coefficient β and the independent variables; Π refers to the multiplicative equivalent of the summation sign (Σ) meaning that *LF* multiplies the values for each case. Thus, the aim is to identify β values producing L_i and P_i values that maximize *LF*.

To avoid the problem of dealing with tiny numbers due to the *multiplication of probabilities, the likelihood function can be converted into a logged likelihood function by taking the natural logarithm of both sides of equation (4.1) and simplifying as follows:*

$$ln\, LF = Y i * lnPi + 1 - Yi * ln1 - Pi \qquad (31)$$

In practice, MLE aims to find those β values that have the greatest likelihood of maximizing the log-likelihood function (Pampel, 2000). Since the likelihood equations are usually non-linear in β, general-purpose iterative methods, such as Newton-Raphson and Fisher Scoring methods, are used for estimating β.

2. Newton-Raphson Method (NRM): The Newton-Raphson method is an iterative procedure for solving non-linear equations, such as those whose solution defines the maximum point of a function. Starting with a guess about the answer, it obtains a second estimate by approximating the function to be maximized by a second-degree polynomial in the region of the initial guess and then determining the location of the polynomial's maximum value. Then it uses another second-degree polynomial to estimate the function in the vicinity of the second guess. The third guess is the position of the function's maximum. The approach creates a succession of guesses in this fashion. When the function is appropriate and the original estimate is reasonable, they converge to the position of the maximum. To determine the value of $\hat{\boldsymbol{\beta}}$ at which the function $L(\boldsymbol{\beta})$ is to be maximised, let $\boldsymbol{u}' = (\partial L(\boldsymbol{\beta})/\beta_1, (\partial L(\boldsymbol{\beta})/\beta_2, \ldots)$; $\boldsymbol{H}$, called the *Hessian matrix*, denote the matrix comprising the entries $h_{ab} = (\partial^2 L(\boldsymbol{\beta})/\partial\beta_a\beta_b)$ and is also known as the *observed information matrix*; $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{u}^{(t)}$ and $\boldsymbol{H}^{(t)}$ be $\boldsymbol{\beta}$, $\boldsymbol{u}$ and $\boldsymbol{H}$ at step t in the iterative process $(t = 0,1,2,\ldots)$. The following relationship, assuming that $\boldsymbol{H}^{(t)}$ is non-singular, is obtained according to Agresti (2002, p.*143):*

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left(\boldsymbol{H}^{(t)}\right)^{-1} \boldsymbol{u}^{(t)} \qquad (32)$$

In the case of a logistic regression model with binary dependent variable Y and one independent variable X, as explained in section 2.2, $\boldsymbol{\beta}$, $\boldsymbol{u}$ and $\boldsymbol{H}$ are given by:

$$\boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \boldsymbol{u} = \begin{bmatrix} \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \frac{e^{\alpha + \beta X_i}}{1 + e^{\alpha + \beta X_i}} \\ \sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} \frac{X_i e^{\alpha + \beta X_i}}{1 + e^{\alpha + \beta X_i}} \end{bmatrix}, \boldsymbol{H} = \begin{bmatrix} -\sum_{i=1}^{n} \frac{e^{\alpha + \beta X_i}}{\left(1 + e^{\alpha + \beta X_i}\right)^2} & -\sum_{i=1}^{n} \frac{X_i e^{\alpha + \beta X_i}}{\left(1 + e^{\alpha + \beta X_i}\right)^2} \\ -\sum_{i=1}^{n} \frac{X_i e^{\alpha + \beta X_i}}{\left(1 + e^{\alpha + \beta X_i}\right)^2} & -\sum_{i=1}^{n} \frac{X_i^2 e^{\alpha + \beta X_i}}{\left(1 + e^{\alpha + \beta X_i}\right)^2} \end{bmatrix}$$

3. Fisher Scoring Method (FSM)

The Fisher Scoring Method (FSM), resembling the Newton-Raphson method (NRM), is an alternative iterative method for solving likelihood equations. The distinction between the two methods is concerned with the *Hessian matrix*. The NRM uses the *Hessian matrix* itself, also called the *observed information matrix*. In contrast, the FSM uses the *expected* v*alue* of this matrix, called the *expected information matrix*.

Let $\boldsymbol{J}^{(t)}$ refer to the approximation t for the maximum likelihood estimate of the expected information matrix, that is, $\boldsymbol{J}^{(t)}$ has elements $-E\,(\partial^2 L(\boldsymbol{\beta}) \,/\, \partial\beta_a\beta_b)$, evaluated at $\boldsymbol{\beta}^{(t)}$. The formula for FSM is given by (Agresti, 2002, p. 145-146):

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left(\boldsymbol{J}^{(t)}\right)^{-1} \boldsymbol{u}^{(t)} \qquad (33)$$

There are various approaches to test the Goodness of Fit (GoF) of a binary logistic regression model which are discussed below:

1. Information Criteria Test (ICT): This is a single statistic for comparing models; better-fitted models have lower values of the same information criterion (Hilbe, 2015). The two widely used ICTs are:

- Akaike Information Criterion (AIC) test defined by the statistic:

$$AIC = -2 \ln L + 2p \tag{34}$$

- Bayesian Information Criterion (BIC) or Schwarz Criterion (SC) defined the test statistic

$$BIC\ (or\ SC) = -2 \ln L + p \ln n \tag{35}$$

Where L is the likelihood function, p is the number of parameters, and n is the sample size.

2. Deviance and Pearson χ^2. In this approach, the data represents the fit of the ideal model possible- the saturated model having a separate parameter for each observation. It is tested whether all parameters that are in the saturated model but not in the estimated model equal zero.

$$\text{Deviance: } D = 2 \sum_{i=1}^{l} \sum_{j=1}^{k} n_{ij} \ln \left(\frac{n_{ij}}{\hat{n}_{ij}} \right) \tag{36}$$

$$\text{Pearson } \chi^2\text{: } \chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{k} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} \tag{37}$$

Where n_{ij} are the observed counts and $\hat{n}_{ij}$ are the counts in the estimated model. Note that D and χ^2 follow chi-square distribution with degrees of freedom equal to the difference between the number of profiles (parameters estimated in the saturated model) and estimated parameters.

However, if the logistic regression model satisfies either of the following conditions:

- Includes continuous variables
- Includes many independent variables
- Independent variables have a considerable number of categories

Note that the Deviance and Pearson χ^2 do not follow χ^2 distribution. In that case, the **Hosmer-Lemeshow (HL)** test is recommended which is calculated as follows:

- Estimating the individual probabilities
- Sorting within response and dividing the set into ten groups with a similar number of observations (k =1,2,3,…,10)
- Using the groups to calculate expected counts and compare with observed counts using Pearson χ^2 statistic, which approximately follows χ^2 distribution with v=10–2=8 degrees of freedom

3. Pseudo-R^2. Several other good of fit statistics have been suggested as analogues of the R^2 statistic often used in linear regression. In these formulae, $L0$ is the likelihood of regression with only an intercept, $L1$ is the likelihood of the model estimated, and n is the sample size.

$$\textbf{McFadden R}^2\textbf{:}\ R^2_{McF} = 1 - \frac{lnL_1}{lnL_0} \tag{38}$$

$$\textbf{Cox-Snell R}^2\textbf{:}\ R^2_{CS} = 1 - \left(\frac{L_0}{L_1}\right)^{2/n} \tag{39}$$

The main drawback of this statistic is that its upper bound is not one, but $1 - L_0^{2/n}$ Also, R^2_{CS} can be normalized using the following transformation by Nagelkerke (1991)

$$\textbf{Normalized Cox-Snell R}^2\textbf{:}\ R^2_N = \frac{R^2_{CS}}{1 - L_0^{2/n}} \tag{40}$$

(Long & Freese, 2014; Pampel, 2000)

2.1.2.5. Ordinal vs Multinomial Logistic Regression model

For this we created the following table that summarises the main characteristics and points out the most relevant characteristics of both model techniques.

Ordinal Logistic Regression	Multinomial Logistic Regression
Logit Model: Ordinal logistic regression is appropriate when the outcome variable is ordinal and has more than two ordered categories. Cumulative Logit Model or Proportional Odds Model is a particular type of model that considers the ordering of categories and	**Logit Model:** When the outcome variable is nominal having more than two unordered categories, multinomial logistic regression is appropriate. Let the model include one outcome variable Y having 3 unordered categories (Y = 0, 1, 2) and one dichotomous independent

Ordinal Logistic Regression	Multinomial Logistic Regression
assumes that the odds ratio is invariant to where the outcome categories are dichotomized. Let the model include one ordinal outcome variable *Y* having 3 ordered categories (*Y* = 0, 1, 2) and one dichotomous independent variable X_1 ($X_1 = 0, 1$). Then there are two ways to dichotomize the outcome: ($Y \geq 1$ *vs.* $Y < 1$; $Y \geq 2$ *vs.* $Y < 2$). With this categorization of *Y*, the logit model is given by two regression equations: $\text{odds}\ (Y \geq 1) = \frac{P\ (Y \geq 1 / X_1)}{P\ (Y < 1 / X_1)} = \alpha_1 + \beta_1 X_1$ $\text{odds}\ (Y \geq 2) = \frac{P\ (Y \geq 2) / X_1}{P\ (Y < 2) / X_1} = \alpha_2 + \beta_1 X_1$ **Conditional Probabilities:** $P\ (Y \geq 1 / X_1) = \frac{1}{1 + e^{-(\alpha_1 + \beta_1 X_1)}}$ $P\ (Y \geq 2 / X_1) = \frac{1}{1 + e^{-(\alpha_2 + \beta_1 X_1)}}$ **Odds Ratio (OR):** $\text{OR}_1 = \frac{[P\ (Y \geq 1 / X = 1) / P\ (Y < 1 / X = 1]}{[P\ (Y \geq 1 / X = 0) / P\ (Y < 1\ / X = 0]} = e^{\beta_1}$ $OR_2 = \frac{[P\ (Y \geq 2 / X = 1) / P\ (Y < 2 / X = 1]}{[P\ (Y \geq 2 / X = 0) / P\ (Y < 2 / X = 0]} = e^{\beta_1}$	variable X_1 (X_1 = 0, 1). Also, let the conditional probabilities be: $P\ (Y = 0 / X_1) = P_0$; $P\ (Y = 1 / X_1) = P_1$; $P\ (Y = 2 / X_1) = P_2$. Two regression equations give the logit model: $ln \left[\frac{P\ (Y=1 / X_1)}{P\ (Y=0 / X_1)}\right] = \alpha_1 + \beta_{11} X_1$ $ln \left[\frac{P\ (Y=2 / X_1)}{P\ (Y=0 / X_1)}\right] = \alpha_2 + \beta_{21} X_1$ **Conditional Probabilities:** The probabilities are given by: $P_0 = \frac{1}{1 + e^{\alpha_1 + \beta_{11} X_1} + e^{\alpha_2 + \beta_{21} X_1}}$ $P_1 = \frac{e^{\alpha_1 + \beta_{11} X_1}}{1 + e^{\alpha_1 + \beta_{11} X_1} + e^{\alpha_2 + \beta_{21} X_1}}$ $P_2 = \frac{e^{\alpha_2 + \beta_{21} X_1}}{1 + e^{\alpha_1 + \beta_{11} X_1} + e^{\alpha_2 + \beta_{21} X_1}}$ **Odds Ratio (OR):** $\text{OR}_1 = \frac{[P\ (Y = 1 \mid X = 1) / \text{P}\ (Y = 0 \mid X = 1]}{[P\ (Y = 1 \mid X = 0) / \text{P}\ (Y = 0 \mid X = 0]} = e^{\beta_{11}}$ $\text{OR}_2 = \frac{[P\ (Y = 2 \mid X = 1) / \text{P}\ (Y = 0 \mid X = 1]}{[P\ (Y = 2 \mid X = 0) / \text{P}\ (Y = 0 \mid X = 0]} = e^{\beta_{21}}$

Source: Author compilation from Kleinbaum and Klein (2010), Agresti (2002).

2.1.2.6. Probit analysis

Probit analysis uses scores from the cumulative standard normal distribution instead of calculating logged odds from the logistic distribution. Logistic regression employs the logistic curve, while probit analysis uses the cumulative standard normal curve. Corresponding to the formula in logistic regression for the logged odds, $L_i = In\ (P_i\ /\ (1 - P_i\ ),$ the formula for probit analysis identifies the *inverse* of the cumulative standard normal distribution. If the cumulative standard normal distribution is represented by Φ, then $P = \Phi\ (Z),$ and $Z = \Phi^{-1}\ (P\ ),$ where Φ^{-1} refers to the inverse of the cumulative standard normal distribution. Although a simple formula cannot represent it, the inverse of the cumulative standard normal distribution transforms probabilities into linear Z scores representing the dependent variable in probit analysis. With probits as the dependent variable, the estimated coefficients show the change in *z* score units of the inverse of the cumulative standard normal distribution rather than the probability change. Even though independent variables have a non-linear relationship to the probabilities, z scores from the probit transformation are linear.

Although the logit curve reaches the floor and ceiling somewhat quicker than the probit curve, the differences are insignificant. With simple formulae, one can convert probabilities into logged odds and vice versa in logistic regression. Probit analysis is more challenging because of the complicated formula for the standard normal curve. This means that programs must apply an arbitrary normalization to set the scale of logit and probit variables, as they have no inherent scaling properties. Using probit analysis, the standard deviation of the error is equal to 1, but using logit analysis, it is around 1.814. Probit and logit coefficients cannot be directly compared because of their different error variances. The logit coefficients will be about 1.8 times greater than the probit coefficients.

It is possible to divide the logit coefficients by that factor to keep the units equivalent. However, logistic regression and probit coefficients will differ because of the modest discrepancies between the logistic and normal curves. Probit analysis does not have the same multiplicative odds coefficients as logistic regression, contributing to its increased popularity. Given the transformed units of the dependent variable, probit coefficients are interpretable in the same way as other regression coefficients. They illustrate the linear and additive change in the probit transformation's z-score units (i.e., the inverse of the cumulative standard normal distribution) when the independent variables are changed by one unit. Perhaps even less evident than logged odds, the cumulative normal distribution's standard units are of little interpretative significance (Pampel, 2000).

2.1.3. Regularisation methods

Regularization techniques are used in machine learning to minimize the variance of estimated parameters and hence the variability of model predictions in an ill-posed problem or overfitting the model. The first scenario of an ill-posed problem pertains to econometric models that display estimation instability. A slight change in the training set might cause huge changes in the model's predictions; this is undesirable, particularly in credit scoring, can lead to an unsuitable model. It may be argued that model stability is a prerequisite for technological robustness in ethical and trustworthy machine learning. The multicollinearity of explanatory variables may cause the ill-posed issue; for example, variables representing credit arrears in multiple tenors, distinct tenors of DPD, and dummy behavioural variables are instances of highly correlated characteristics. The second situation of overfitting the model occurs when a model performs well on the training set. However, the prediction error increases considerably on different data sets. The model is considered to be overfitted to the initial data set (training set) and hence unable to infer on other data sets (Kaszyński et al., 2020). The inclusion of too many independent variables on a training set may lead to a factitiously high prediction power on training sets but not on test sets (Melkumova & Shatskikh, 2017). The three types of regularisation methods, ridge regression, LASSO regression, and elastic net regularisation, are discussed below.

2.1.3.1 Ridge Regression

Ridge regression is a regularization technique used to estimate parameters when the independent variables display near multicollinearity, is a method of biased estimation with biased estimated parameters (i.e. $E\left[\hat{\beta}\right] \neq \beta$). However, this allows for the reduction of the variance of estimated parameters. Hence, model streamlining may reduce the model predictions and parameter values variance while incurring low estimator bias. That trade-off (reduced variance and low bias) is desirable from the commercial point of view unless it significantly impairs the model's interpretability. The ridge regression parameters $\hat{\boldsymbol{\beta}}$ are defined as an optimal solution for the following inequality-constrained quadratic programming problem (Melkumova & Shatskikh, 2017):

$$\arg\min_{\beta_i}\left[\sum\nolimits_{i=1}^{N} log\ (\ 1\ +\ exp\ -\sum\nolimits_{j}\beta_j\, x_{i,j}y_i\ ))\ +\ t\sum\nolimits_{j}\beta_j^2\right] \qquad (41)$$

Here t ≥ 0 is the penalty term or the tuning parameter. Based on the minimisation of the prediction error on the validation set, a model with a specific tuning parameter (target model) is chosen for a given set of *t* parameters.

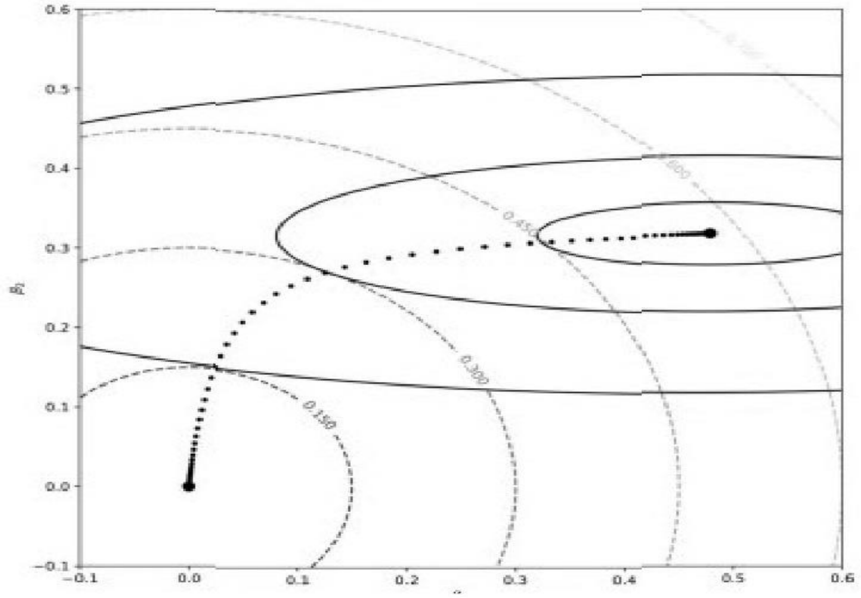


Figure 2.1: Contour curve of ridge regression loss function

[Source: Reproduced with permission from Kaszyński et al. (2020, p. 99)]

2.1.3.2 Lasso Regression

In the case of Lasso (Least Absolute Shrinkage and Selection Operator) regression, the regression parameters $\hat{\boldsymbol{\beta}}$ are defined as an optimal solution for the following constrained problem (Yang & Wen, 2018).

$$\arg\min_{\beta_i}\left[\sum\nolimits_{i\,=\,1}^{N} log\;(\,1\,+\,exp\,-\sum\nolimits_{j}\beta_j\,x_{i,j}y_i\,))\,+\,t\sum\nolimits_{j}|\,\beta_j^2\,|\right] \tag{42}$$

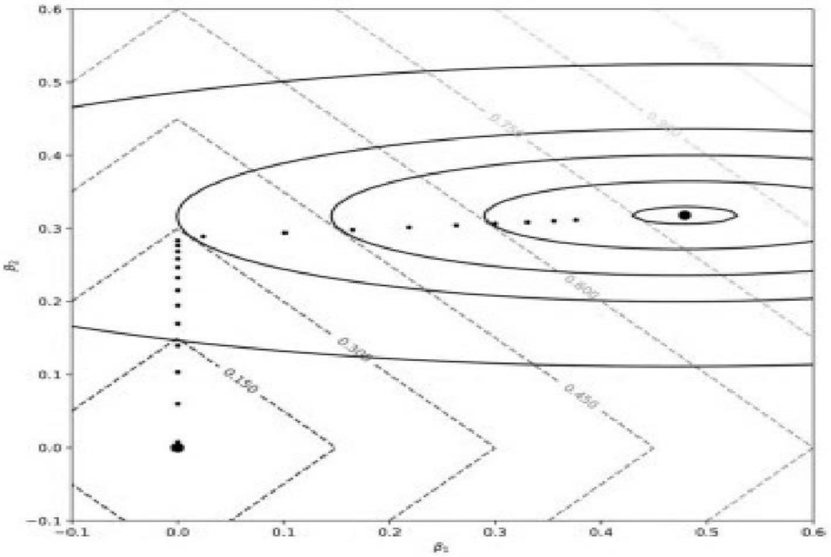
Where t ≥ 0 is the penalty term or the tuning parameter.

Akin to ridge regression, lasso regression sets a limit on the norm of model parameters; in the case of lasso regression, it is the *l*1 norm or urban norm. Even though these two regularisation methods have similar model structures (Adding an additional penalty component imposed on the loss function), they may be differentiated as follows:

- Lasso regression allows the exclusion of individual parameters by making them equal to zero (β_i = 0). In contrast, ridge regression allows the values of the model parameters to be reduced to minimal values that are not equal to zero.

- The lasso regression makes parameter interpretation simpler than ridge regression.
- In the case of ridge regression, the change in the sign for a specific parameter may be observed, whereas, in the case of lasso regression, the model parameters change more linearly with the tuning parameter.
- Ridge regression imposes a larger penalty than lasso regression.

Since Lasso regression can make the coefficients equal to zero, this method is better when dealing with multicollinearity and selecting relevant variables.



2.1.3.3 Elastic net

The *elastic net* regularization combines both the penalties terms used in ridge and lasso regressions. Zhou et al. (2015) show that the elastic net may be reduced to the linear Support Vector Machine (SVM); for each setup of the elastic net (concerning the parametrization of penalties) in the case of binary classification problem, the solution of the elastic net is similar to the hyper-plane solution of linear SVM, thus enabling the use of highly optimized SVM solvers and Graphics Processing Unit (GPU) acceleration for elastic net problems. The logistic regression loss function with elastic net regularization assumes the following form:

$$\arg\min_{\beta_i}\left[\sum_{i=1}^{N} log\,(1 + exp - \sum_j \beta_j\, x_{i,j}y_i\,)) + \lambda\,((1-\alpha)\,t\sum_j |\beta_j| + \alpha\,(\sum_j \beta_j^2\,)^{\frac{1}{2}}\,)\right] \tag{43}$$

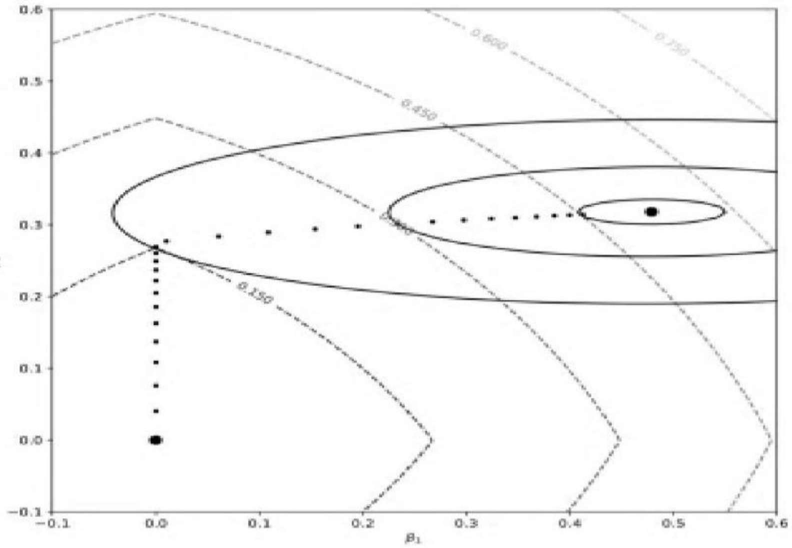


Figure 2.3: Contour curve of the elastic net regression loss function

[Source: Reprinted with permission from Kaszyński et al. (2020, p. 102)]

2.2 Non-Parametric models used in credit scoring

Non-parametric models require a few or no assumptions concerning the relationship between dependent variable and explanatory variables, distribution of the variables or the errors include decision trees, random forest, extreme gradient boosting (XGBoosting), and neural network discussed below.

2.2.1 Decision Trees

A decision tree is an algorithm of supervised learning that can be interpreted in the form of a graphical tool, with a branch or root-like structure of boxes and lines used to show possible turns of events that may or may not be controllable (Anderson, 2007). When the dependent variable is nominal and accepts multiple values indicating object classes, trees are often utilized in classification issues. For example, in the context of credit scoring, a classification tree is a machine learning technique that classifies a collection of observations into subgroups based on attribute value tests. The division's objective is to create subsets. The most often used techniques for building classification trees minimize the measure of diversity in the resulting subgroups (Hastie et al., 2009). Initially, efforts to derive decision trees were made by trial and error. According to Thomas et al., (2017), Breiman and Friedman separately proposed utilizing analytical techniques to derive the rule set in 1973. However, it was not until 1984 that they collaborated with Olshen and Stone to create Trees (CART) algorithms and C4.5 algorithms (Breiman, 1996; Quinlan, 1993). The credit scoring literature points to Makowski and Coffman as one of the earliest implementations of decision trees in credit scoring analysis. Another subset of algorithms are those that split the set using statistical testing. This category includes methods such as the CHAID algorithm, which uses the $\chi 2$ statistics (Kass, 1980) and permutation tests-based algorithms (Hastie et al., 2009).

The tree structure represents the recursive divisions. The nodes and leaves signify the obtained subsets, while the tree branches illustrate the division rules. When the dependent variable is nominal and accepts multiple values indicating object classes, trees are often utilized in classification issues. The explanatory variables establish the divisions of observation sets. The tree nodes indicate the results of tests on the variable values. The root node is the first node that encompasses all observations. The branches correlate to the findings of many tests. The leaves display the category labels for the observations. Each tree represents a collection of classification rules, and each leaf represents a distinct classification rule. A classification rule is a statement of the form: IF THEN ELSE statements.

Classification rules are a helpful technique to describe information in an easily interpretable manner. They suggest the appropriate options to make when the criteria are satisfied. In the instance of classification, the rule is referred to as classification and is as follows: IF conditions are met, THEN a category is created.

The splits in the example depicted in Figure 2.4 are calculated from the top down. The *root node* is at the top of the tree, each succeeding level is referred to as a *child node*, and the *terminal nodes* are at the bottom. There will be two or more splits each time, and the number of levels will vary based on the branch. When completed, the values of the terminal nodes may be utilized as estimations (scores) or as a grouping tool. For a binary result, the value is a probability, such as P(Bad), and any branches with probabilities more significant than a preset cut-off, such as greater than the average P(Bad), are classified as Bad (Anderson, 2007).

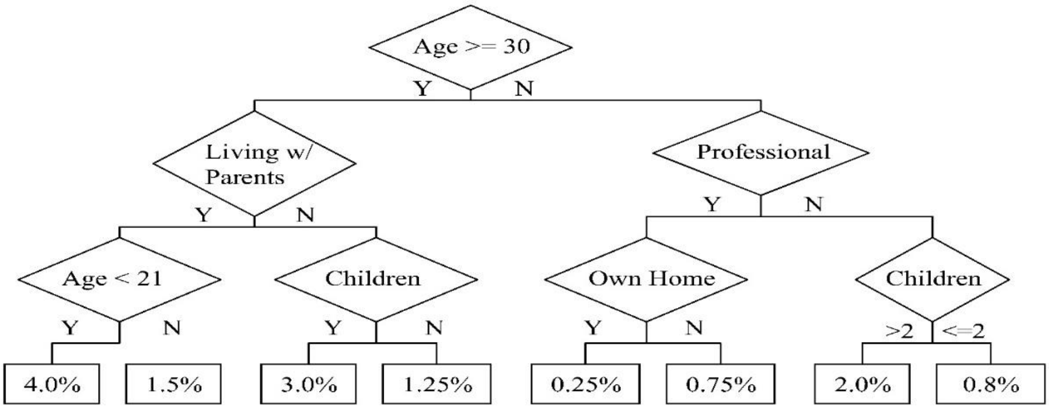


Figure 2.4: Decision tree

[Source: Reprinted with permission from Anderson (2007, p. 173)]

The fundamental approach used is called a Recursive Partitioning Algorithm (RPA) which outlines the methods of finding the branches through repeated attempts to locate the optimal split. The following criteria define the RPA process:

- **binning**, ascertaining how predictors should be binned.
- **splitting**, choosing which characteristic to utilize.
- **stopping**, when to stop the creation of new sub-nodes.
- **pruning**, how to remove nodes to prevent overfitting.
- **assignment**, how to categorise each node as good or bad (Anderson, 2007).

Splitting Criteria

Selecting the optimum test for dividing the set entails deciding on the characteristic to test and the available test outcomes. Each node of the same tree may have a unique test set accessible. Additionally, tests conducted at distinct nodes may be based on the same variable yet provide distinct sets of findings. There are various approaches for the splitting criteria, such as entropy and information gain, information gain ratio, Gini index, Twoing criteria, and CHAID (Kaszyński et al., 2021).

Binning and variable report

Binning is a method for discretising a numerical continuous variable by grouping its values to a series of bins. This method is useful for dealing with outliers by bucketing them in the lowest or highest intervals of the distribution with less extreme values. In this method, outliers become identical to other values in the distribution's tail. Binning may also resolve issues caused by skewness. Binning may be unsupervised (bins are created solely based on the variable's distribution) or supervised (bins are produced using external information, e.g. target variable). In practice, variable binning should be performed for logistic regression; for decision trees, feature binning should be performed inside the algorithm (Kaszyński et al., 2021; Mansour & Schain, 2001; Quinlan, 1993)

Stopping criteria

Stopping criteria determine whether or not the tree-building process should be halted. A tree that has been over-expanded is overfitted to the training set. As a result, its quality is often substantially worse on the test set. Preventing a tree from overfitting is accomplished via the use of halting criteria and tree trimming. The halting condition may prevent overfitting by prohibiting further division of the node. The criteria for halting may also be technological in character. These technical requirements include the following:

➢ set homogeneity, i.e. when a node contains only data from a single class;

➢ absence of accessible tests, i.e. when all observations in a node have the same values for the explanatory variables.

The following criteria are used to prevent a tree from becoming overfit:

★ maximum depth of the tree – when a tree reaches the maximum depth specified in the process parameters, the tree's construction is complete;

★ minimum node size – a node will not be split if its child nodes are less than the specified minimum node size;

★ minimum leaf size – a node will not be split if the obtained leaves are less than the specified minimum leaf size; the node will become a leaf;

★ Minimum leaf purity - if node's purity does not fall below a certain threshold, the node becomes a leaf;

★ minimal increase in purity – if the increase in purity is less than the provided amount, the node will not be split:

If node X purity is measured by entropy, then $\triangle P(t, X) = g_t\ (X)$ (Kaszyński et al., 2020).

2.2.1.3 Advantages of classification trees

- A classification tree's structure enables easy comprehension of the resulting relationships and categorization criteria, even those without sophisticated statistical expertise. The tree's shape illustrates precisely which variables, how they interact, and in what sequence they impact the assignment of data to certain categories. The relevance of specific variables (which indicates how much it helps reduce impurity in the derived subsets) within the tree structure and hence within the classification process may be determined.
- Classification trees are an excellent way to visualise the non-linear connections in the data. They may be constructed using any variable type - real, ordinal, or nominal. Moreover, they do not need any specific variable processing prior to model construction, but they may be treated in the same way as other approaches; for example, binary variables or weighted averages can substitute nominal variables. Furthermore, these methods perform well when dealing with outliers, missing numbers, and errors in data. There is no need for the pre-selection of variables.
- A tree-building approach based on recursive divisions enables the selection of the most relevant variables for the classification process. They are invariant to numerical variable monotone transformations. However, monotonicity restrictions may be implemented, which increases the transparency and compliance of derived models with accessible information (Kaszyński et al., 2020).

2.2.1.4 Disadvantages of classification trees

- A minimal change in the value of a single variable may often totally alter the categorization result. On the other hand, a minor modification to the data set may change the tree's overall structure.
- The interpretability of decision trees rapidly decreases as the number of nodes increases (Kaszyński et al., 2020).

4.4.4 Model

4.4.5 Assess

4.5 Comparison of models

5. Conclusions

6. References

References

Abdou, H. A., & Pointon, J. (2011). Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature. *Intelligent Systems in Accounting, Finance and Management*, *18*(2–3), 59–88. https://doi.org/10.1002/ISAF.325

Anderson, Raymond. (2007). *The credit scoring toolkit : theory and practice for retail credit risk management and decision automation*. Oxford University Press.

Breiman, L. (1996). Technical Note: Some Properties of Splitting Criteria. *Machine Learning 1996 24:1*, *24*(1), 41–47. https://doi.org/10.1023/A:1018094028462

Demajo, L. M., Vella, V., & Dingli, A. (2020). *Explainable AI for Interpretable Credit Scoring*. 185–203. https://doi.org/10.5121/csit.2020.101516

EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679, (2016) (testimony of European Parliament). https://gdpr-info.eu/

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188. https://doi.org/10.1111/J.1469-1809.1936.TB02137.X

Handy, C. (2002). What is a Business for? *Harvard Business Review*, 9. https://ssrn.com/abstract=932676

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). Prentice Hall.

Kass, G. v. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, *29*(2), 119. https://doi.org/10.2307/2986296

Kaszyński, D., Kamiński, B., Szapiro, T., Kwiatkowski, M., Przanowski, K., Zając, S., Opiński, Ł., Wrzosek, M., Siuta, K., Cerazy, K., Chlebus, M., Kłosok, M., Biecek, P., Kraiński, Ł., & Nosarzewski, A. (2021). *Credit Scoring in Context of Interpretable Machine Learning* (D. Kaszyński, B. Kamiński, & T. Szapiro, Eds.). SGH Publishing House. https://doi.org/10.5281/ZENODO.4692106

Kennedy, K. (2013). Credit Scoring Using Machine Learning. *Doctoral*. https://doi.org/https://doi.org/10.21427/D7NC7J

Knutson, M. (2020). *Credit Scoring Approaches Guidelines*. The World Bank Group. https://thedocs.worldbank.org/en/doc/935891585869698451-0130022020/CREDIT-SCORING-APPROACHES-GUIDELINES-FINAL-WEB

Langdon, R. J., Yousefi, P. D., Relton, C. L., & Suderman, M. J. (1992). The Degradation of the Scorecard over the Business Cycle. *IMA Journal of Mathematics Applied in Business and Industry*, *4*, 497–509. https://doi.org/10.2/JQUERY.MIN.JS

Lauer, J. (2017). *Creditworthy : a history of consumer surveillance and financial identity in America*. Columbia University Press.

Lee, J., & Hogarth, J. M. (2018). The Price of Money: Consumers' Understanding of APRs and Contract Interest Rates: *Https://Doi.Org/10.1177/074391569901800108*, *18*(1), 66–76. https://doi.org/10.1177/074391569901800108

Lewis, E. (1992). *An Introduction to Credit Scoring* (Reprint edition). Athena Press.

MacDonald, S. B., & Gastmann, A. L. (2017). A History of Credit & Power in the Western World. *A History of Credit and Power in the Western World*, 1–314. https://doi.org/10.4324/9781315083513

Mansour, Y., & Schain, M. (2001). Learning with Maximum-Entropy Distributions. *Machine Learning 2001 45:2*, *45*(2), 123–145. https://doi.org/10.1023/A:1010950718922

Martin, V., & Evzen, K. (2006). Credit-Scoring Methods. *Czech Journal of Economics and Finance (Finance a Uver)*, *56*(3–4), 152–167. https://EconPapers.repec.org/RePEc:fau:fauart:v:56:y:2006:i:3-4:p:152-167

Melkumova, L. E., & Shatskikh, S. Y. (2017). Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering*, *201*, 746–755. https://doi.org/10.1016/J.PROENG.2017.09.615

Mohanty, N., John, A. L. S., Manmatha, R., & Rath, T. M. (2013). Shape-Based Image Classification and Retrieval. *Handbook of Statistics*, *31*, 249–267. https://doi.org/10.1016/B978-0-444-53859-8.00010-2

Onay, C., & Öztürk, E. (2018). A review of credit scoring research in the age of Big Data. *Journal of Financial Regulation and Compliance*, *26*(3), 382–405. https://doi.org/10.1108/JFRC-06-2017-0054

Puertas, R., Olmeda, I., & Bonilla, M. (2003). *Parametric and non-parametric models in the credit scoring problems*. https://www.researchgate.net/publication/289978389_Parametric_and_non-parametric_models_in_the_credit_scoring_problems

Quinlan, J. R. (John R. (1993). *C4.5 : programs for machine learning*. Morgan Kaufmann Publishers. https://books.google.com/books/about/C4_5.html?id=b3ujBQAAQBAJ

Siddiqi, N. (2017). Intelligent Credit Scoring. *Intelligent Credit Scoring*. https://doi.org/10.1002/9781119282396

Siddiqi, Naeem. (2013). *Credit risk scorecards : developing and implementing intelligent credit scoring*. https://books.google.com/books/about/Credit_Risk_Scorecards.html?hl=fr&id=SEbCeN3-kEUC

Thomas, L., Crook, J., & Edelman, D. (2017). Credit Scoring and Its Applications, Second Edition. *Credit Scoring and Its Applications, Second Edition*. https://doi.org/10.1137/1.9781611974560

World Bank Group. (2019). Disruptive Technologies in the Credit Information Sharing Industry. *Disruptive Technologies in the Credit Information Sharing Industry*. https://doi.org/10.1596/31714

Yang, X., & Wen, W. (2018). Ridge and Lasso Regression Models for Cross-Version Defect Prediction. *IEEE Transactions on Reliability*, *67*(3), 885–896. https://doi.org/10.1109/TR.2018.2847353

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). John Wiley & Sons.

Browne, I. (1997). Explaining the Black-white gap in labor force participation among women heading households. *American Sociological Review, 62*(2), 236-252. https://doi.org/10.2307/2657302

Cramer, J. (2003). The Origins of Logistic Regression. *Tinbergen Institute Discussion Paper*.

https://doi.org/10.2139/ssrn.360300

DeMaris, A. (1993). Odds versus probabilities in Logit equations: A reply to Roncek. *Social Forces, 71*(4), 1057-1065. https://doi.org/10.2307/2580130

DeMaris, A., Teachman, J., & Morgan, S. P. (1990). Interpreting logistic regression results: A critical commentary. *Journal of Marriage and the Family, 52*(1), 271-277.

https://doi.org/10.2307/352857

Hosmer, D. W., Lemeshow, S., & Sturdivant, R.X. (2013). *Applied Logistic Regression* (3rd ed.). John Wiley & Sons.

Kaufman, R. L. (1996). Comparing effects in dichotomous logistic regression: A variety of standardized coefficients. *Social Science Quarterly*, 77, 90-109.

Kleinbaum, D. G., & Klein, M. (2010). *Logistic Regression: A Self-Learning Text* (3rd ed.). Springer Science & Business Media.

Long, J. S., & Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. SAGE.

Long, J. S., & Freese, J. (2014). *Regression models for categorical dependent variables using Stata* (3rd ed.). Stata Press.

Pampel, F. C. (2020). *Logistic Regression: A Primer*. SAGE Publications.

Roncek, D. W. (1993). When will they ever learn that first derivatives identify the effects of continuous independent variables or "Officer, you can't give me a ticket, I wasn't speeding for an entire hour". *Social Forces, 71*(4), 1067-1078. https://doi.org/10.2307/2580131