UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI "M.FANNO"

DIPARTIMENTO DI MATEMATICA "TULLIO LEVI-CIVITA"

CORSO DI LAUREA MAGISTRALE IN ECONOMICS AND FINANCE

TESI DI LAUREA

"CREDIT RISK ANALYSIS WITH MACHINE LEARNING TECHNIQUES IN THE PSD2 FRAMEWORK: THE BUDDYBANK CASE STUDY"

RELATORE:

CH.MO PROF. CLAUDIO FONTANA

LAUREANDO: TOMMASO CANZIAN

MATRICOLA N. 1179640

ANNO ACCADEMICO 2019 – 2020

Il candidato dichiara che il presente lavoro è originale e non è già stato sottoposto, in tutto o in parte, per il conseguimento di un titolo accademico in altre Università italiane o straniere.
Il candidato dichiara altresì che tutti i materiali utilizzati durante la preparazione dell'elaborato sono stati indicati nel testo e nella sezione "Riferimenti bibliografici" e che le eventuali citazioni testuali sono individuabili attraverso l'esplicito richiamo alla pubblicazione originale.

The candidate declares that the present work is original and has not already been submitted, totally or in part, for the purposes of attaining an academic degree in other Italian or foreign universities. The candidate also declares that all the materials used during the preparation of the thesis have been explicitly indicated in the text and in the section "Bibliographical references" and that any textual citations can be identified through an explicit reference to the original publication.

Firma dello studente

____

Contents

Introduction

Challenges in Financial industry

Traditional banking currently is in front of a crossroad: continue with the *modus operandi* that has characterized it for decades or undertake a challenging process of digitalization that will change permanently the way of doing business. In fact now more than ever banks are facing a crucial period, not because they are addressing any financial crisis, rather because they are in the middle of a transformation that concerns the banking sector. Companies based on cutting-edge technology are entering into the financial industry, offering really competitive services and products, often more personalized and cheaper than the existing ones in the market. This innovation is called FinTech, which combines the efficiency, flexibility and scalability of technology with the knowledge of the financial world. Recently also giant companies in the tech industry, such Apple or Amazon, have stepped into, giving rise to a new phenomenon called BigTech and it is matter of time before a tech company like Facebook gets a full banking license in order to operate without restrictions.

Research Background

Currently, one of the most affected banking services is lending, in 2016 according to Cambridge Centre for Alternative Finance $284 billion have been globally provided. FinTech companies are deploying groundbreaking advanced analytics technologies such Artificial Intelligence and Machine Learning to assess credit risk and provide financing solutions alternative to traditional banking. This phenomenon is increasingly becoming a research topic in credit scoring literature. The paper *How do Machine Learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm* (Gambacorta et al, 2019) compares the predictive power of credit scoring models based on Machine Learning techniques with that of traditional default models. Authors argue that model based on advanced techniques and non-traditional data is better able of predict default rate during shock to credit supply, because it can more effectively mine the non-linear relationship between target variable and features. In this regard it worth mentioning another paper, *Consumer credit-risk models via machine-learning algorithms* (E.A.

Khandani et al, 2010) where the Classification And Regression Trees (CART) method is used to construct nonlinear nonparametric forecast models very accurate in predicting credit events 3-12 months in advance. This thesis offers a similar approach, modern machine learning algorithms such Random Forest (an advanced method based on CART) and Neural Networks are compared with a more traditional model like Logistic Regression in a task of Default customer classification.

Research Question

Credit risk, or in general risk management, is one of the main activities of a bank that strongly affects its stability and financial sustainability. Accuracy in the prediction is a significant prerogative for a credit scoring model because any wrong decision entails present and future costs for the bank, for example keep granting financing to an undeserving customer. Data analytics can provide solutions to deal with this phenomenon and manage credit card risk, deciding when and how much to cut individual-account credit lines. Therefore the main question that this thesis aims to answer is:

- *Can Supervised Machine Learning techniques help in credit card risk decision-making with respect to traditional models?*
  To address this question the thesis will focus on models' performances on the same dataset to investigate if machine learning models can identify patterns or relationships that escape traditional approaches.

Methodology

The research method follows this structure:

- *Chapter 1*
  Focus on traditional credit risk model (Linear Discriminant Analysis) and reference to the new frontier of credit scoring (Inductive models) from a theoretical point of view.

Then a brief guidance of regulatory changes about expected credit losses inside the IRFS 9 framework is provided.

- *Chapter 2*
  Introduction to Artificial Intelligence and Machine Learning with a presentation of the main ideas underlying these concepts. Deep dive in their current applications in the financial industry and future opportunities.

- *Chapter 3*
  Overview of FinTech industry, literature review about which are the possible drivers of this innovation. Analysis of its evolution and identification of potential benefits and risks for traditional banks in Credit market.

- *Chapter 4*
  Modelling application and evaluation:
    - Presentation of Buddybank dataset
    - Pre-processing phase in order to get tidy data
    - Feature selection and Models implementation
    - Assessing performances in terms of AUC
    - Results

Chapter 1

Credit Risk Models

According to A. Resti and A. Sironi credit risk means "the possibility that an unexpected change in a counterparty's creditworthiness may generate a corresponding unexpected change in the market value of the associated credit exposure"[1]. Reading this definition we can immediately come up with the three main concepts that compose the aforesaid notion:

1. Default risk and migration risk – credit risk is not only linked to the possibility of the counterparty's default, but the worsening of its creditworthiness condition should also be taken into consideration. With this in mind, credit risk measurement and management should be founded not on a simplistic binomial distribution (default vs non-default) but rather on a discrete or continuous distribution, where the default is an extreme event.

2. Risk as unexpected event – this assumption is fundamental in credit risk management because otherwise we would deal with a predictable factor/event and so the probabilistic computation and modelling could be not necessary.

3. Credit exposure – it refers to the total amount of credit that the lender grants to the borrower and its magnitude measures the extent to which the lender is exposed to the risk of loss.

Afterwards, the next logical step is to underline the distinction between Expected loss and Unexpected loss.

[1] A. Resti, A. Sironi, *Risk Management and Shareholders' Value in Banking*, Wiley, 2007

a) Expected loss (EL)

It corresponds to the mean value of the probability distribution of future losses, and for this reason it represent the average size of the risk. It is estimated ex ante by the lender which on the other side in order to ward its position charges an appropriate spread to the interest rate.

To estimate the expected loss we need three parameters:

i. Exposure at default (EAD) : how much of the original amount loaned will be outstanding at the time of default.
ii. Probability of default (PD) : how likely is that borrower will default.
iii. Loss given default (LGD) : what percentage of the amount owed will the lender lose in case of default.

$$EL = EAD \ \times PD \ \times LGD$$

b) Unexpected loss (UL)

It can be defined as the variability of the loss around its mean value.

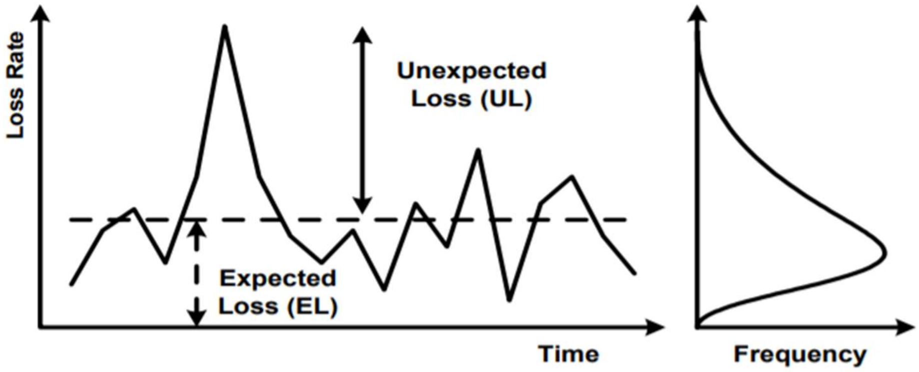


Figure 1.1 Expected and Unexpected Loss[2]

Where the frequency in Figure 1.1 describes the likelihood of losses of a certain magnitude.

[2] Basel Committee on Banking Supervision, *An Explanatory Note on the Basel II IRB Risk Weight Functions*, Bank for International Settlements (BIS), 2005, p. 2

1.1 Credit scoring Models

Risk management becomes critical in this framework, unexpected losses will occur but a bank does not know in advance the timing or the severity and charging such interest rates to potentially cover the entire credit portfolio is infeasible. So the main function of the bank capital is to provide a guarantee to debtholders against unexpected losses.

The two factors a bank can operate on are the valuation of the Expected Loss and the Capital required by financial regulator. In this research I will focus on the first element, especially on the probability of default analysis, the idea behind this choice is that if a bank increases the predictive power of its risk model, accordingly improving the accuracy of the outcome, the loss rate variability could decrease and therefore the unexpected loss, with a gain in terms of capital requirement.

Generally in the literature we have three categories of credit-scoring models:

1- Linear discriminant analysis
2- Regression models
3- Inductive models

1.1.1 Linear Discriminant Analysis

Linear discriminant analysis was studied firstly by Fisher[3] in 1936 and basically it tries to identify the variables which make possible to discriminate between positive or negative instances (default , no default). The model is based on a discriminant function.

[3] R.A. Fisher, *The Use of Multiple Measurements in Taxonomic Problems*, Annals of Eugenics, 1936

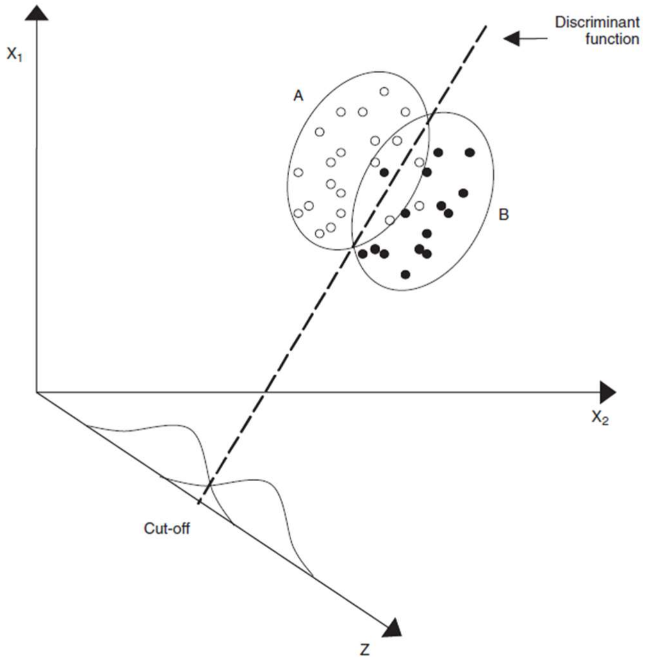


Figure 1.2 Discriminant Analysis[4]

In this simplified scenario we are describing two groups of observations, A the healthy customers (no default), B the insolvent ones (default), using *n* features. The score obtained with these *n* variables is shown on the z axis. In linear discriminant analysis the z score is a linear combination of the independent variables (in Figure 1.2 X_1 and X_2), so generally speaking:

$$z = \sum_{j=1}^{n} \gamma_j \, X_j$$

The coefficients γ_j are chosen in order to maximize the distance between the means (z_A and z_B, also called centroids) of the two groups and we can find them doing:

$$\gamma = \sum^{-1} (X_A - X_B)$$

Where X_A and X_B are vectors containing the mean values of the independent variables for the two classes of observations and Σ is the variance-covariance matrix for the independent variables valid for both classes:

$$\Sigma = \frac{n_A - 1}{n_A + n_B - 2} \Sigma_A + \frac{n_B - 1}{n_A + n_B - 2} \Sigma_B$$

[4] A. Resti, A. Sironi, *Op.cit.,* p. 288

Doing this we are basically "clustering" our observations, so we can set a cut-off threshold α, a point halfway between the two centroids, below which any loan's request is rejected because considered too risky:

$$\alpha = \frac{1}{2}\gamma'(X_A + X_B)$$

Where γ' is the transpose vector of γ.

The index used to measure the discriminant capacity of a model is the Wilks' Lambda:

$$\Lambda = \frac{\sum_{i\in A}(z_i - z_A)^2 + \sum_{i\in B}(z_i - z_B)^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2}$$

Where $\bar{z}$ is the mean of z_i in the entire sample of observations. If the model is effective and it can clearly separate the two sub-groups the Wilks' Lambda is approximately 0, otherwise it will be close to 1.

We can switch from the z score to the probability of default estimation of any observation through an exponential transformation of linear functions of X[5]:

$$PD = p(B|X_i) = \frac{1}{1 + \frac{1-\pi_B}{\pi_B}e^{z_i - \propto}}$$

Where π_B is the *a priori* probability of default, which represents the average probability of default (across the banks' loan portfolios) given the current market conditions.

Discriminant analysis has been the dominant methodology, in terms of publications regarding credit scoring[6], however, it has been criticized for the restrictive assumptions on which the results are based:

[5] E.I. Altman et al., *Application of Classification Techniques in Business, Banking and Finance*, JAI Press, 1981

[6] E.I. Altman, A. Saunders, *Credit risk measurement: Developments over the last 20 years*, Journal of Banking & Finance, 1997

- Normality of the variables, within each class we assume the X variables are distributed with a multivariate normal distribution:

$$X_A \sim N(\mu_A, \Sigma_A)$$

  This assumption enables a straightforward computation of the maximum-likelihood estimator of the covariance matrix.

- Same variance-covariance matrix of the variables in each class (default and no default), homoscedasticity:

$$\Sigma_A = \Sigma_B$$

This make it unemployable in the reality, pushing banks to consider other less stringent models[7].

1.1.2 Regression models

In the regression model we try to estimate the coefficients of the independent variables that lead to the identification of the target variable (default probability) usually through the ordinary least squares:

$$y_i = \alpha + \sum_{j=1}^{n} \beta_j x_{i,j} + \varepsilon_i$$

But with this setting we face a huge problem, the y_i estimated can go outside the range 0-1 making meaningless the result. So, in order to come up with this drawback we introduce the logit model where the linear relationship is adjusted through an exponential transformation which I will discuss in more detail in Chapter 4. This feature allows us to generate a result limited to the interval [0,1] fitting perfectly with the concept of probability.

[7] R.A. Eisenbeis, *Pitfalls in the Application of Discriminant Analysis in Business, Finance, and Economics*, Journal of Finance, 1977

1.1.3 Inductive models

The models listed so far share a common ground: they attempt to find the fundamental relationship that can explain the economic/financial reasons underlying a loan's default. In other words, these models rely on structural characteristics which explain the state of health of an applicant and the variables' selection reflects *a priori* choices based on economic reasoning. The modelling process, in this structural approach, always starts with assumptions made by analysts, while the inductive models operate differently. In fact they use a purely inductive process, starting from a dataset if a pattern is found they leverage it to define relationship between target variable and features without assessing any prior assumption.

Structural models are definitely more transparent, whose logic can be fully understood, but if it is an advantage on one side, on the other it could be also a drawback, in fact these models can be learned and neutralize. While the inductive models, as neural networks, are black boxes hard to decipher and therefore penalized by the regulators. Currently several studies prove how more advanced models as Artificial Neural Networks can outperform traditional methods as Discriminant Analysis or Logistic Regression[8], given also the robustness of the ANN to the stringent assumptions of statistical models such LDA. The paper by H. Lu[9] instead demonstrates that an hybrid model which uses the Logistic Regression in the pre-processing phase for the feature selection (in order to reduce dimensions) and the Neural Networks can be very effective in credit scoring. Furthermore, Machine Learning models can very easily move from the particular (individual PD forecasting) to the general (macroeconomic forecast of credit risk in lending business) being therefore capable of generate indicators of deterioration in consumer creditworthiness and measure of systemic risk. In fact Khandani et al. applying ML to time-series delinquency and default rate in the pre-crisis period (2007-2009) prove how these non-linear nonparametric models are more adaptive and efficient in picking up cycle's dynamics than traditional credit scores[10]. The inductive approach can be exploited by the banks to discover relationships in the data and provide useful insights to the analysts without being necessarily implemented in the official risk assessments.

This whole topic will be discussed more deeply in the Machine Learning section of Chapter 4, when we will see how a Neural Networks (NN) algorithm works.

[8] M.D. Odom, R. Sharda, *A Neural Network Model for Bankruptcy Prediction*, International Joint Conference on Neural Networks, 1990

[9] H. Lu, *Credit Scoring Model Hybridizing Artificial Intelligence with Logistic Regression*, Journal of Networks, 2013

[10] A.E. Khandani et al., *Consumer credit-risk models via machine-learning algorithms*, Journal of Banking & Finance, 2010

Research studies on using NNs for credit risk started in 1990, and still now are very active. It is in fact reasonable claim that in finance nonlinear approach would be superior to a non linear approach. In this regard A. F. Atiya in his paper[11] argues that there are saturation effects in the relationship between explanatory variables, such as financial ratios, and response variable, such as default prediction, reporting the following example, “if the earnings/total assets changes say by an amount of 0.2, from -0.1 to 0.1, it would have a far larger effect (on the prediction of default) than it would if that ratio changes from say 1.0 to 1.2”[12]. This example implies that linearity can sometimes underestimate important aspects taken instead into account by more complex models as Neural Networks.

One of the first applications of Neural Networks to the bankruptcy prediction, by Odom and Sharda[13], compares the predictive power of NN and Multivariate Discriminant Analysis approach using financial ratios. The discriminant analysis method, which was by far the most widely used method for bankruptcy analysis at the time, obtained a correct classification rate of 86.84% for the bankrupt firms in the training sample, therefore caution should be exercised in assessing the robustness of the model. On the same subsample, the trained neural network correctly predicted all the bankrupt firms, with an accuracy rate of 100%. Comparing the models with the hold-out method, which consists in assess the validity of the model on unseen data, the Neural Networks significantly outperforms the Discriminant Analysis.

1.2 Uses of credit scoring models

Therefore credit scoring models have two main purposes:

1- Separate healthy from risky loans;
2- Estimate the probability of default of each loan;

In the first case we set a threshold below which the credit application is rejected, and all the loans above this cut-off point are considered equally reliable (in machine learning terms this is called Classification).

[11] A.F. Atiya, Bankruptcy Prediction for Credit Risk Using Neural Networks: A Survey and New Results, IEEE Transactions on Neural Networks, 2001
[12] Idem, p. 2
[13] M.D. Odom, R. Sharda, *Op.cit.*

The second use has a more granular approach because it consists in the risk assessment of every borrower, regardless of the loan's status, and a probability of default is assigned to each borrower. But sometimes the starting hypothesis behind these models are unrealistic, thus banks have compensated this issue grouping customers with similar scores and observing in the following years the percentage of default within each score range and then assuming this percentage as PD estimate for the customers belonging to a given class (this is called actuarial approach).

1.3 New Regulation: IRFS 9

The financial crisis of 2007-2008 highlighted the issue of early recognition of credit losses, therefore the International Accounting Standards Board introduced the International Financial Reporting Standards (IFRS) 9, which is an accounting standard that imposes on banks to detect increased credit risk loans timely and take the necessary measures, building corresponding provisions based on expected credit losses (ECL)[14]. When a deterioration in the credit quality is observed, a reclassification of the loan takes place, but what is important to remark is that loss identification is no more founded on the occurrence of a triggering event (for example the borrower loss of employment) as in the past directives, but rather on a more forward-looking approach[15]. The reason of this new view is that loan pricing may not reflect the risk because of market conditions and then for a matter of economic capital, during good times decisions on adequate capital are more efficient than during stressed periods. Not timely recognition of loan losses considerably raises the negative impact of recession on banking lending[16]. The same thing is supported by the Financial Stability Forum which noted that an earlier identification of loan losses could have dampened cyclical moves in the 2007-2008 financial crisis. However finding a threshold for exactly defining an increase in credit risk is challenging, because a conservative choice could result in a high ECL calculation volatility, because a potential credit downturn would cause the downgrade of several loans, despite the temporary nature of the event. While a lenient approach if on one side creates a more stable calculation on the other could determine delays in the recognition of dangerous loans. A solution could be assessing the

[14] B.H. Cohen, G.A. Edwards Jr, *The new era of expected credit loss provisioning*, BIS, 2017

[15] L. Ewanchuk, C. Frei, *Recent Regulation in Credit Risk Management: A Statistical Framework*, MDPI, 2019

[16] A. Beatty, S. Liao, *Do delays in expected loss recognition affect banks' willingness to lend?*, Journal of Accounting and Economics, 2011

increased credit risk on a collective basis of financial instruments, this would ensure that expected credit losses are identified in case of significant increase of risk even if evidence at individual level is not yet available.

In this framework of increasing complexity in assessing credit risk and evident incompleteness of traditional techniques, AI and machine learning can play a crucial role in risk management, supporting decision makers in difficult situations where past experiences cannot help or when not all the information are available. A proper and regulated use of machine learning can reduce the chance of future financial disasters caused by incorrect judgements.

Chapter 2

Artificial Intelligence and Finance

2.1 Artificial Intelligence

Artificial intelligence recently has become a buzzword due to the improvement in computing power of modern computers and the increase in available data. Application of AI has concerned many fields, from self-driving car to finance and the revenues derived from AI-related product and service are expected to exceed three trillions of US Dollars by 2024[17].

But what really is Artificial Intelligence?

Artificial Intelligence (AI) initially was defined as a "branch of Computer Science that is concerned with the automation of intelligent behavior"[18], where intelligent behavior, according to Turing[19] refers to the ability to achieve human-level performance in all cognitive tasks, sufficient to fool an interrogator. It is popular the Imitation Game proposed by Turing immediately after the World War II in order to assess the ability to think of a machine and which has been recognized for years as the most reliable way to define a machine intelligent, although it has recently been reformulated and revised[20]. A more up to date definition of AI is proposed by the European Commission, "AI systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from these data and deciding the best actions to take to achieve the given goal"[21].

Artificial Intelligence in general is a very broad field, of which machine learning or ML is a sub-category. Machine learning may be defined as the ability to learn from data without being explicitly programmed and with limited human intervention. So this is in contrast with the rules-based algorithm which was the traditional practice in the computer science field, where human

[17] Statista , "Revenues from The Artificial Intelligence (AI) Market Worldwide From 2015 to 2024"

[18] G.F. Luger, W.A. Stubblefield, *Artificial Intelligence: structures and strategies for complex problem solving*, Addison Wesley Publishing Company, 1997

[19] A.M. Turing, *Computing Machinery and Intelligence*, Mind, 1950

[20] T.C.M. Tse et al., *The AI Republic*, Lioncrest Publishing, 2019

[21] European Commission, *Ethics Guidelines for Trustworthy AI*, 2018

programmer explicitly determines what decisions are being taken under particular states of the world.

If with classical algorithms we need to come up first with the best solution and then we program the computer to apply it faster and more efficiently, with machine learning it is up to the algorithm find the best solution given a prior training process where the algorithm autonomously makes hypotheses and selects the optimal one. The key point in machine learning lies in its flexibility, in fact ML tries to find patterns in large amount of data not constrained by linear relationships, often imposed in traditional economic and financial theory. It is noteworthy to say that ML merely identifies correlations but it cannot determining causality. Machine learning is a combination of different fields, as the Venn diagram shows (Figure 2.1), it is a convergence of mathematical/statistical knowledge with computer science skills.

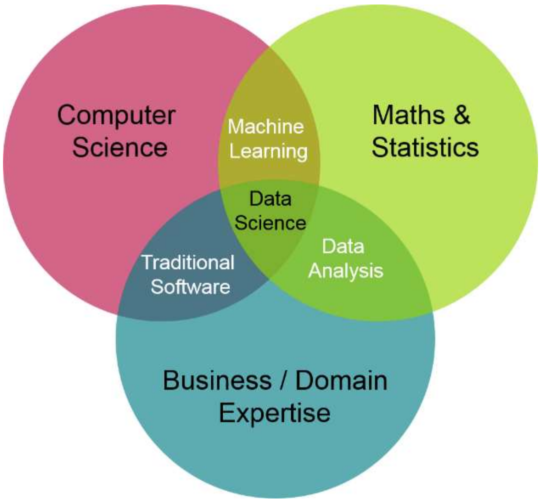


Figure 2.1 Venn Diagram of Data Science

2.2 How Machine Learning works

Doing machine learning means gathering and cleaning the data, understanding which algorithm could be optimal in that particular task and feed the algorithm with the data so that it can learn relationships (here we are in the phase of abstraction, where the data has been transformed into an abstract form aimed to summarize the original information and thereby new relationships before ignored can emerge). The learning process is still not complete until the model is able to make use of the abstracted knowledge for unseen data. Now the generalization phase takes over and it involves the reduction of the hypothetical theories inferred from the data (about the relationship between inputs and target variable) into a reasonable number through a heuristic-based thinking. The heuristics used by machine learning algorithms, decisive in reducing computational time, can run into erroneous conclusions (bias) during this process, because the assumptions made can produce errors. But in a scenario with no assumptions the algorithm would have no better performance than a random selection. This theory, formalized by Macready and Wolpert in 1997, is called the "no free lunch" theorem[22] and it states that "any two optimization algorithms are equivalent when their performance is averaged across all possible problems".

In his paper[23] Mitchell defines bias as "any basis for choosing one generalization over another, other than strict consistency with the instances", pointing out the fundamental importance of bias in the learning process, in fact without bias a classifier model could not predict any instance that is not identical to the training instance because it would not be able to provide a basis for generalization.

Bias implicitly refers to an error, more precisely to the error due to bias, which is the difference between the expected prediction of our model and the actual value. While the error due to variance is related to the variability of a model prediction for a given data point, so variance is a measure to assess the generalizability of the model. Doing machine learning also involves optimize both variance and bias: bias-variance tradeoff.

[22] D.H. Wolpert, W.G. Macready, *No Free Lunch Theorems for Optimization*, IEEE Transactions on Evolutionary Computation, 1997

[23] T.M. Mitchell, *The Need for Biases in Learning Generalizations*, 1980

2.2.1 Bias - Variance tradeoff

Consider a supervised learning algorithm A and an unknown target function f to be learned given the input space X to the real numbers ℝ. Let $S = \{(x, f(x) + \varepsilon \mid x \in X\}$ be a training sample with noise ε. During the generalization process the algorithm A will output an hypothesis $A(S) = \hat{f}$, so for a given test point x_0, the predicted value is $\hat{f}(x_0)$. Suppose that during the training phase we repeatedly draw training samples $S_1, \ldots, S_l$ each of size m and therefore the algorithm A formulates l hypotheses $\widehat{f_{S_1}}, \widehat{f_{S_2}}, \ldots, \widehat{f_{S_l}}$ where $\widehat{f_{S_\iota}}$ represents the hypothesis $A(S_i)$, we can obtain the expected predicted value of x_0 :

$$\bar{\hat{f}}(x_0) = \frac{1}{l}\sum_{i=1}^{l} \widehat{f_{S_\iota}}(x_0)$$

The Bias of algorithm A at point x_0 is:

$$Bias(A, x_0) = \bar{\hat{f}}(x_0) - f(x_0)$$

The Variance of algorithm A at point x_0 is:

$$Variance(A, x_0) = E[\left(\hat{f}(x_0) - \bar{\hat{f}}(x_0)\right)^2]$$

The variance captures the variation from one training set to another, which can derives from variation in training sample, random noise ε, or random behavior in the learning algorithm.

In order to evaluate the performance of a model the most commonly used measure is the mean squared error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}\,(x_i))^2$$

Where $\hat{f}(x_i)$ is the prediction for the ith observation while y_i is the actual value.

If we consider the MSE for a given point x_0 (with actual value y_0) we can decompose it in three parts: the variance of $\hat{f}(x_0)$, the square bias of $\hat{f}(x_0)$ and the variance of irreducible error terms ε:

$$Expected\ MSE = E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0) + [Bias\left(\hat{f}(x_0)\right)]^2 + Var(\varepsilon)$$

In fact:

$$E\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = E[y^2] + E\left[\hat{f}(x_0)^2\right] - 2E\left[y\hat{f}(x_0)\right]$$

$$= Var(y) + E^2[y] + Var\left(\hat{f}(x_0)\right) + E^2\left[\hat{f}(x_0)\right] - 2E\left[y\hat{f}(x_0)\right]$$

$$= Var(y) + E^2[y] + Var\left(\hat{f}(x_0)\right) + E^2\left[\hat{f}(x_0)\right] - 2E[\varepsilon]E\left[\hat{f}(x_0)\right] - 2E\left[f(x)\hat{f}(x_0)\right]$$

A fundamental assumption is the independence between ε and the train and test samples.

$$= Var(y) + E^2[y] + Var\left(\hat{f}(x_0)\right) + E^2\left[\hat{f}(x_0)\right] - 2E[\varepsilon]E\left[\hat{f}(x_0)\right] - 2E\left[f(x)\hat{f}(x_0)\right] - 2Cov(f(x), \hat{f}(x_0))$$

$$= Var\left(f(x)\right) + Var(\varepsilon) + E^2\left[\hat{f}(x_0)\right] - 2E[\varepsilon]E\left[\hat{f}(x_0)\right] - 2E\left[f(x)\hat{f}(x_0)\right] - 2Cov\left(f(x), \hat{f}(x_0)\right)$$

$$= Var\left(f(x) - \hat{f}(x_0)\right) + Var(\varepsilon) + E^2[y] + E^2\left[\hat{f}(x_0)\right] - 2E[\varepsilon]E\left[\hat{f}(x_0)\right] - 2E\left[f(x)\hat{f}(x_0)\right]$$

$$= Var\left(f(x) - \hat{f}(x_0)\right) + Var(\varepsilon) + E^2[f(x_0)] + E^2[\varepsilon] + 2E[\varepsilon]E[f(x)] + E^2\left[\hat{f}(x_0)\right] - 2E[\varepsilon]E\left[\hat{f}(x_0)\right] - 2E\left[f(x)\hat{f}(x_0)\right]$$

$$= Var\left(f(x) - \hat{f}(x_0)\right) + Var(\varepsilon) + \left(E[f(x)] - E\left[\hat{f}(x_0)\right]\right)^2 + E^2[\varepsilon] + 2E[\varepsilon]E[f(x)] - 2E[\varepsilon]E\left[\hat{f}(x_0)\right]$$

Now we assume that the noise ε has zero mean,

$$= Var\left(f(x) - \hat{f}x_0\right) + \left(E[f(x)] - E[\hat{f}(x_0)]\right)^2 + Var(\varepsilon)$$

As stated before, the goal is to reduce both bias and variance, that instead have an “inverse” relationship, in order to find the optimal model complexity (Figure 2.2).

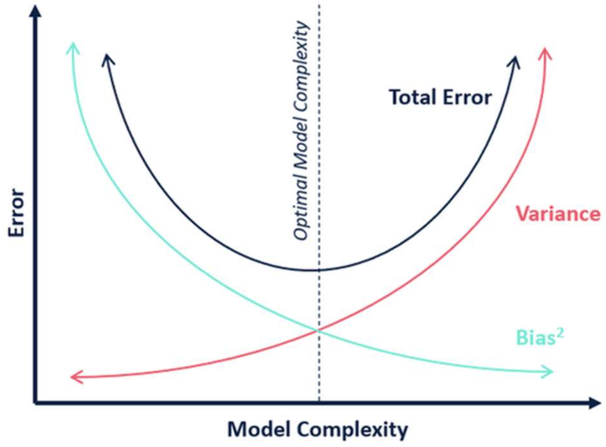


Figure 2.2 Bias-Variance Tradeoff[24]

The two extreme cases of bias/variance tradeoff are underfitting and overfitting. Underfitting occurs when a model is unable to capture the underlying pattern of the data, and it is characterized by high bias and low variance. This case can happen when we have a small sample of data which is not representative of the population, or when we build a too simplistic model, maybe trying to catch non-linear relationships with a linear model. Potential solutions could be: using more features to increase predictive power or trying a more complicated model. Overfitting instead is when the model is too flexible and it captures also the noise in the data, so it fits too well the pattern of the training data with a consistent problem in generalization. To overcome this risk we can use fewer features in order to decrease variance, or train the model on more data.

2.2.2 Types of Machine Learning

We can divide machine learning algorithms in three main groups[25]:

- Supervised learning (mainly used to build predictive models)
- Unsupervised learning (mainly used to construct descriptive models)
- Reinforcement learning (mainly used in interactive environment such as self-driving car)

[24] S. Ozdemir, *Principles of Data Science*, Packt, 2017, p. 245

[25] A. Ng, *Machine Learning Yearning*, 2018

Supervised learning basically finds associations between features and a target variable, it tries to fit a model that relates the response to the predictors, with two principal purposes: accurately predicting the response variable for future observations, so prediction, or understanding more clearly the relationship between response and predictors, so inference. With supervised machine learning labeled data act as past experience for the model and we not only try to predict new instances, we also can be interested in understanding the relationships that bind our data.

There are two types of supervised learning models: regression and classification; usually we tend to identify the former with problems with a quantitative response while we tend to refer to problems with a qualitative response as classification. The border is blurred when we have binary prediction model.

Whenever labeled data are missing we talk about unsupervised learning, which uses the set of predictors to accomplish different tasks such as dimensionality reduction, condensing variable together, or discover patterns / similar behaviors across data when no apparent structure exists, such as clustering.

A considerable advantage of this kind of learning is that we do not need labels, so it is easier to get the data, but on the other side of course we lose the predictive power and we cannot monitor how well we are doing. Sometimes the unsupervised learning is used as exploratory analysis also in predictive analytics problems, in order to identify groups or clusters.

In reinforcement learning algorithms are rewarded (positively or negatively) for the decisions that they take, so they keep acting with the purpose of maximize their reward function. The reinforcement learning use is widespread in AI-assisted game and in robotics (especially in the self-automated machinery), its main drawback is that the learning process can be very articulated, it can employs many attempts before realizing these actions have a negative reward.

2.3 Overview of Machine Learning in financial services

Machine learning has wide-ranging applications in the financial industry, it allows to generate analytical insights, develop new products and services and reduce market frictions, all for the benefit of consumers that obtain more tailored and cheaper products.

Many financial companies have already adopted this technology, according to the last report by Bank of England published on October 2019[26], almost 70% of financial institutions in UK have live machine learning applications in use (mainly the firms in the banking/insurance industry). Moreover Bank of England in its report states that this number is expected to significantly grow, in fact who is not deploying ML today is going to in coming years and the median number of live applications is intended to increase.

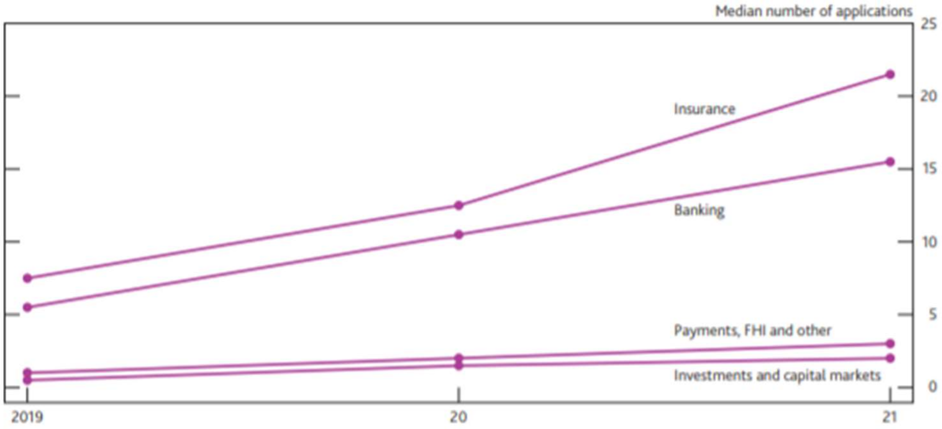


Figure 2.3 Evolution in ML applications[27]

Considering the business areas with a higher rate of application we see at the first place back-office functions such as risk management and compliance which include anti-money laundering, credit scoring and fraud detection activities. However, recently also front-office areas have experienced an increasing use of machine learning, for instance customer management as well as sales and trading.

In any case, for one firm out of two machine learning is a strategic priority, with a dedicated plan for research, development and deployment, in fact only 25% of ML use cases are

[26] Bank of England, *Machine learning in UK financial services*, Financial Conduct Authority, 2019

[27] Idem, p. 9

implemented by third-party providers[28]. Firms also sometimes rely on third-parties for IT platforms and infrastructure such as cloud computing, which is more typical for non-bank firms. In addition to these services, it is not uncommon that companies use data collected by third-parties from different industries and combine them with existing data to gain new insights.

For what concern benefits from ML and future expectations, the areas mentioned in the report as more involved are: fraud detection and anti-money laundering, followed by operational efficiency and products personalization.

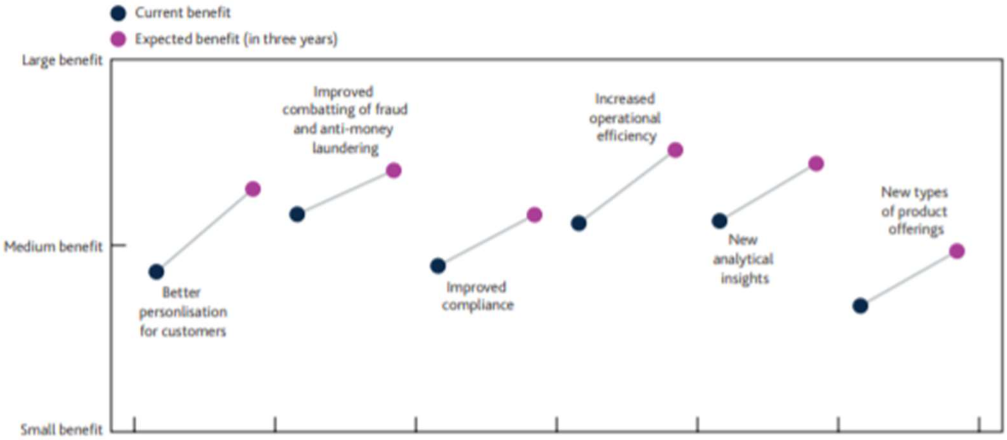


Figure 2.4 Current and Expected Benefits[29]

Fraud detection, and security in general, is a hot topic in finance, due to the increasing numbers of transactions, and ML is excellent in detecting fraud given the nature of the event (a series of suspicious activities), moreover with the huge number of cross data now is even easier. Process automation is also a promising application of machine learning, technology in fact allows to replace manual work improving efficiency, enabling the companies to optimize costs and customer experience.

We have also practical examples from the financial industry:

- JP Morgan has developed a platform, Contract Intelligence (COiN) that through Natural Language Processing can extract essential data from legal documents.

[28] Idem

[29] Idem, p. 16

- Bank of America introduced Erica, a virtual assistant recognized as one of the world's prominent financial service innovation, that leverages cognitive messaging and predictive analytics.
- Citibank has deployed machine learning to implement a sophisticated anti-fraud system for online and mobile banking.
- US Bank applies their effort to developing conversational interfaces and chatbots to improve the customer service.

If on one side machine learning produces benefits, on the other side it can amplify existing risk, because narrowing the human judgment and accelerating the operations can expose the bank. The main uncertainty is represented by the lack of ML model explicability, but also biases in data and algorithms or inadequate control of this technology (a good illustration is the case of Knight Capital in 2012 with their stock algo-trading automation resulting in a loss of $440 million in less than one hour). While potential constraints that can get in the way of deploying ML are: legacy systems principally, this is particularly true for well-established firms in banking and insurance where the *modus operandi* is consolidated and the path dependence is strong (past practices are hard to dispose of), difficulty of integrating ML into existing business processes, because the transition could be not so smooth, and the regulations, which are aimed to maintain financial stability and protect consumers, have a wary attitude towards this new blurred technologies, given that some algorithms are still black boxes. Especially deep learning can entails regulatory compliance issues in demonstrating model validity, due to the model's complexity and low interpretability.

Chapter 3

FinTech Innovations

The word FinTech stands for financial technology and it is used to define these new firms which are based on cutting-edge technologies to improve and make more efficient the delivery and use of financial services. In the last years we witnessed the launch of many startups some of which then became “unicorns” (company valued over $1 billion), worth mentioning Monzo, Revolut and N26, who openly challenged the traditional banking industry. Some paradigms are changed, now interconnection and speed are indispensable values in a service, because people want to have immediate access to a product. The innovations introduced by the fintech firms are not only related to the products and services offered, rather they are part of a new ecosystem populated by heterogeneous agents. A striking example is Apple, who recently launched its own credit card (in partnership with Goldman Sachs) and that is not only a business diversification statement, it is the sign of a new concept of doing business, more integrated and tailored to a person life. If Bank of America does not perceive it as a threat and it merely judges the product features, misunderstanding the disruption, it could make the same mistake of Nokia when in 2007 Apple launched the first iPhone radically mutating the interaction between human and personal phone.

A disrupted area in the financial landscape is the retail lending, where fintech platforms using big data, alternative data and complex machine learning algorithms are able to evaluate borrowers’ credit risk in a more effective and profitable way (the FICO score is even not considered in the creditworthiness determination). This allows consumers with short credit history, that therefore would not satisfy the traditional lending requirements, to get access to loans without compromising the portfolio quality of the fintech firm. A brilliant case of fintech consumer lending is LendingClub, which is the largest fintech lender for peer to peer loans (more than $50 billion financed). It is a platform where potential borrower and investors meet, after the risk assessing by an algorithm, the grade is attributed to the loan applicant and the investors are free to finance the loan to yield the corresponding interest rate. At first glance is presumable thinking that only the rejected from the traditional lending apply for a loan in LendingClub, making the portfolio risky, but data suggests the opposite. In fact this firms leveraging the huge amount of data they annually store due to the numerous requests and exploiting alternative data sources (ignored by traditional banks) they can train more effectively the algorithm making possible an efficient "cream skimming", that is serving only valuable clients. As Figure 3.1 shows, from 2012 to 2019 an improving in the LendingClub portfolio occurred with a gradual disappearance of the worst rated loans.

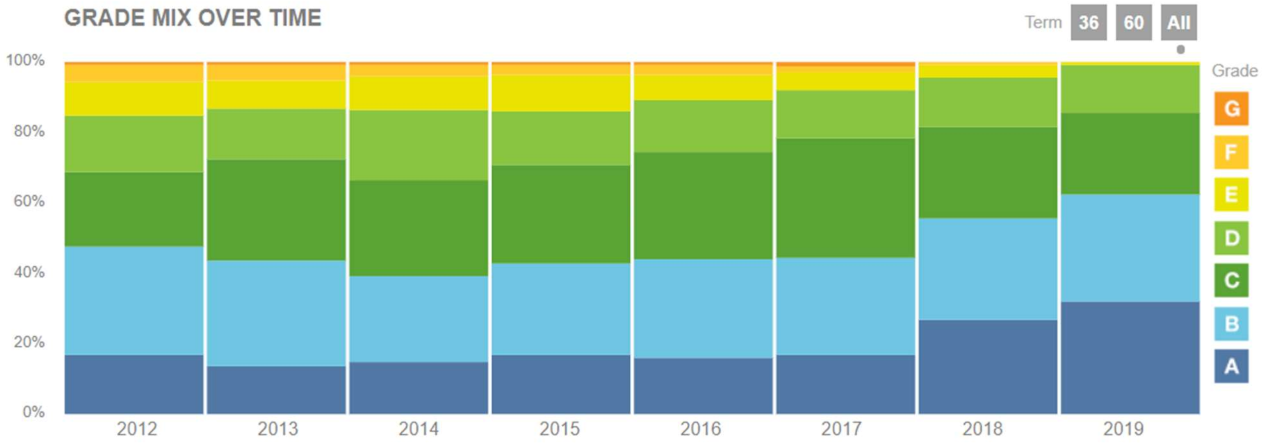


Figure 3.1 Data from LendingClub.com

The gap between the fintech firms and traditional banks is becoming ever wider, if we analyze the evolution of the correlation between FICO score (traditional measure of credit risk) and the rating grade assigned by LendingClub (see Figure 3.2). This discrepancy kept growing in the recent years, pointing that LendingClub has increasingly used non-traditional alternative data in the risk assessment process.

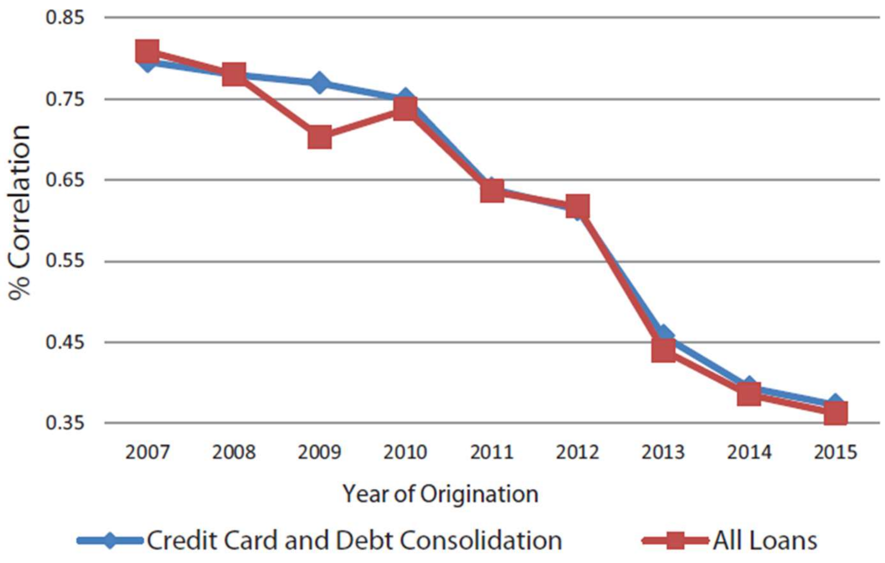


Figure 3.2 Correlation FICO and LandingClub score[30]

[30] J. Jagtiani, C. Lemieux, *The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform*, Wiley, 2019, p. 8

According to a research by Jagtiani and Lemieux[31] the distribution of consumer loans made by LendingClub in the United States varies with the degree of bank branching activities, in fact the declining in bank branches it is a factor that favors the diffusion of LendingClub. Moreover, the study reveals that also the number of bank branches *per capita* (100,000 people) in a county inversely impact the amount of newly originated loans. Therefore LendingClub seems to be filling a potential credit gap (from a geographical point of view) remarking a distinctive benefit of FinTech: the financial inclusion, try to extend services and products to those ignored by traditional channels.

3.1 Machine Learning and Credit risk

The ingredients of this innovative lending formula can be summarized in two elements: the alternative information gathered and exploited by the fintech firms and also the use of new technologies such as Machine Learning algorithms. In addition to a different and more effective approach than the traditional one, fintech lenders process mortgage applications about 20% faster without a higher default rate. Then considering also the fact that fintech credit is usually uncollateralized (big difference with traditional banks), the use of big data is particularly relevant.

In the research[32] conducted by L. Gambacorta et al. they try to assess whether fintech credit scoring models based on machine learning and big data (Model 1) are better in predicting borrowers' defaults than linear models based on traditional (Model 2) and non-traditional (Model 3) data obtained through customers' mobile phone apps and their activity on e-commerce platform.

- Model 1: based on FinTech credit score, where machine learning has been used
- Model 2: based on traditional credit card information using logistic regression
- Model 3: based on traditional and non-traditional information using logistic regression

[31] J. Jagtiani, C. Lemieux, *Do FinTech Lenders Penetrate Areas That Are Underserved by Traditional Banks?*, Journal of Economics and Business, 2018
[32] L. Gambacorta et al., *How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm*, BIS, 2019

It results that the model based on fintech credit score has the highest Pseudo R^2, followed by the third model. This means that, in the considered sample, these new models who lean on machine learning algorithms and big data are better than traditional models that use bank-type information in predicting default rates.

Then the authors analyze the performance of the models in the event of an exogenous shock, given that the machine learning technology is sometimes reputed to be effective only when the relationship between inputs and outputs remains the same, and in financial application, situations of stability are infrequent. They used an unexpected regulatory change as proxy for an exogenous shock and it seems that the Model I outperformed the others, implying the reliability of machine learning models also in dynamic environments.

Focusing on the discriminatory power of each model the authors show the gap between Model I (based on the credit score obtained using machine learning with big data) and Model II (traditional bank model) decomposing it in two parts, the lighter one represents the value added by non-traditional information while the darker one is the gain obtained from machine learning, the dotted red line is the exogenous shock event.

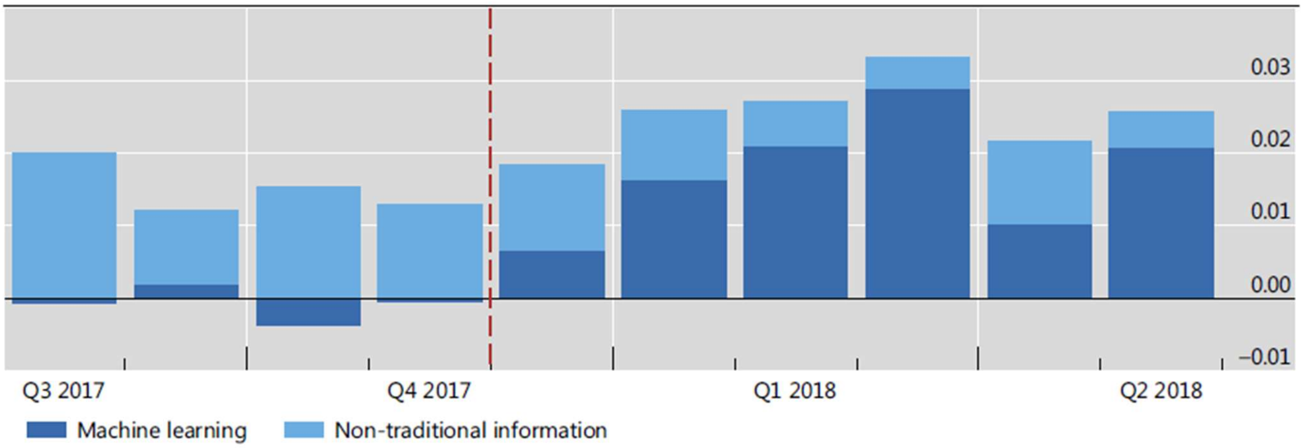


Figure 3.3 Machine Learning value added[33]

The value added by Machine Learning has been calculated comparing the prediction accuracy of Model 1 and Model 3, both trained on the same set of information, but the first one using a ML algorithm while the second one a Logistic regression. In the same way the value added by non-traditional information (using Model 2 and 3) has been computed. In the y axis is

[33] L. Gambacorta et al., *Op.cit.*, p. 16

represented the gap between the models in terms of "area under the curve", which is a common performance measure, higher is the value better is the prediction, so a positive gap implies a better quality of a model with respect to the other.

As can be easily noticed, the contribution of ML is very relevant after the shock, maybe due to the fact that machine learning can extract richer information from the features during period of stress, the non-linearity of the model surely helps in this.

Another interesting aspect considered in the paper is about credit scoring and bank-customer relationship. It is highlighted that the comparative advantage of the fintech model over the traditional one increases for low levels of the bank-customer relationship, while when the relationship becomes stronger the differences decrease. This suggests that machine learning and big data approaches can help in situation of asymmetric information, when the lender does not know the credit history of the customer.

3.2 Credit FinTech markets across the world

According to estimates by the Cambridge Centre for Alternative Finance in 2016 have been granted $284 billion of FinTech credit across the world, from the $11 of 2013. However the growth of FinTech credit (CrediTech) has been quite heterogenous among the different jurisdictions.

In absolute terms, China represented the larger market in 2016, followed by United States and UK, while considering from a *per capita* point of view is remarkable the role played by smaller economies such as Estonia or New Zealand. After an initial boost over the three years period 2013-2016, the data show a slowdown in several of the major jurisdictions, China above all (Figure 3.4 right-hand panel).

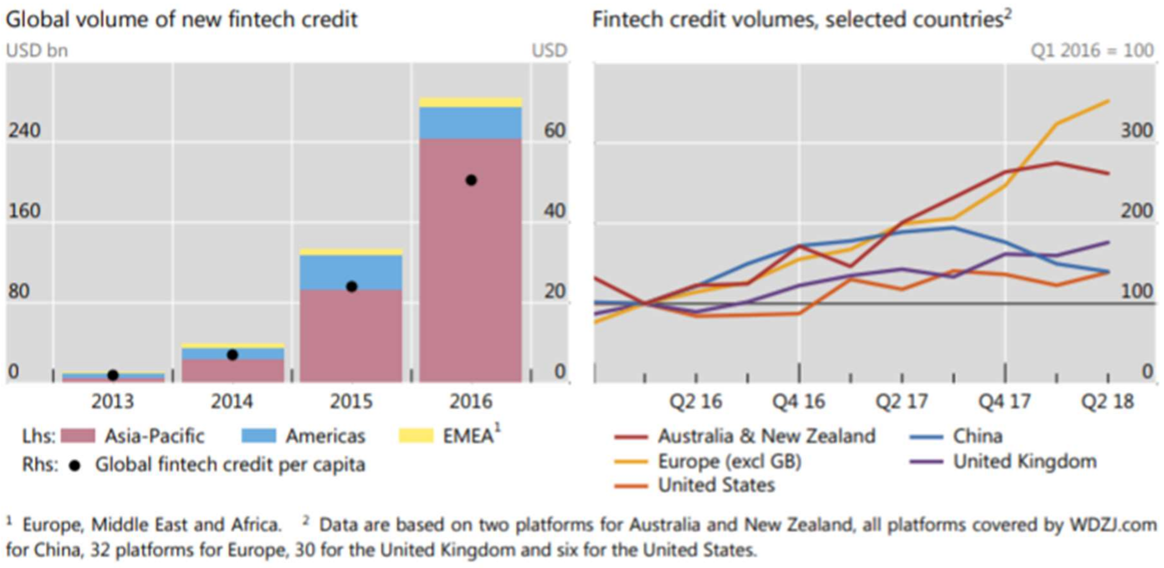


Figure 3.4 Dynamics of FinTech credit around the world[34]

Figure 3.4 highlights how FinTech credit's size differs across countries, implying that multiple factors impact the pace at which the CrediTech market grows.

In order to analyze the drivers of FinTech credit is worth mentioning the paper by S. Claessens et al[35]. They identify country-specific factors such as economic and financial development, the quality of legal institutions or the degree of competition in lending market, a softer competition in banking system and higher margins can facilitate the rise of FinTech companies. An other important factor pointed out in the research is the regulatory stringency, based on the study made by G.B. Navaretti et al[36] which constructed this index using different indicators from the World Bank's Bank Regulation and Supervision Survey to measure the sensitivity of the regulatory system to bank risk-taking.

Results from the bivariate regression implemented by the authors, considering GDP *per capita*, treated as a proxy for economic development, and regulatory stringency as explanatory variables of FinTech credit *per capita* are plotted in the panels of Figure 3.5

[34] S. Claessens et al., *Fintech credit markets around the world: size, drivers and policy issues*, BIS, 2018, p. 34
[35] Idem
[36] G.B. Navaretti et al., *FinTech and Banks: Friends or Foes?*, European Economy – Banks, Regulation, and the Real Sector, 2018

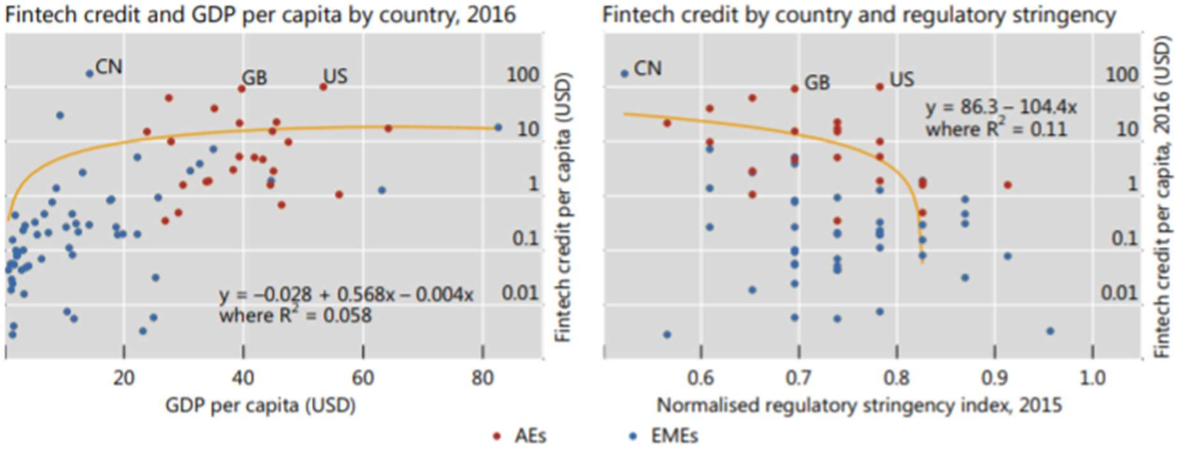


Figure 3.5 CrediTech correlation with economic development and regulatory stringency[37]

Where AEs are advanced economies and EMEs are emerging economies.

FinTech credit has a non-linear positive relationship with the stage of development of a country (the effects become less important at higher levels of development) and a negative relationship with Regulatory stringency (which is normalized to get a value between zero and one).

In order to better understand the relative importance of the different potential drivers, a multivariate regression analysis is implemented for the same sample of 63 economies for 2016:

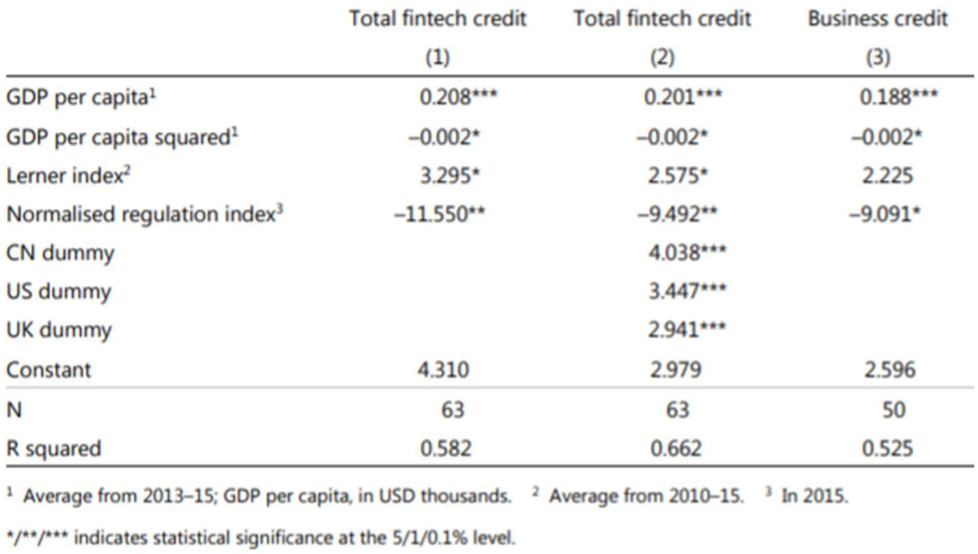
$$c_i = \alpha_i + \beta_1 y_i + \beta_2 {y_i}^2 + \gamma LI_i + \delta RS_i + \varepsilon_i$$

Where c_i is the volume of FinTech credit *per capita* in economy *i* in 2016; y_i is the log of GDP *per capita* in economy *i*, the variable ${y_i}^2$ captures possible non-linearity in the relationship; LI_i is the Lerner Index[38] of banking sector in economy *i* and RS_i is the regulatory stringency index for economy *i* (a higher value indicates a more stringent regulation).

From Figure 3.6 column (1) we can observe the same results presented previously and additionally the significant positive coefficient of the Lerner Index implies that FinTech is more active in less competitive banking sector.

[37] S. Claessens et al., *Op. cit.*, p. 37

[38] A.P. Lerner, *The Concept of Monopoly and the Measurement of Monopoly Power*, The Review of Economic Studies, 1934



	Total fintech credit	Total fintech credit	Business credit
	(1)	(2)	(3)
GDP per capita^{1}	0.208***	0.201***	0.188***
GDP per capita squared^{1}	–0.002*	–0.002*	–0.002*
Lerner index^{2}	3.295*	2.575*	2.225
Normalised regulation index^{3}	–11.550**	–9.492**	–9.091*
CN dummy		4.038***	
US dummy		3.447***	
UK dummy		2.941***	
Constant	4.310	2.979	2.596
N	63	63	50
R squared	0.582	0.662	0.525

^{1} Average from 2013–15; GDP per capita, in USD thousands. ^{2} Average from 2010–15. ^{3} In 2015.

*/**/*** indicates statistical significance at the 5/1/0.1% level.

Figure 3.6 Results from Multivariate regression[39]

It is interesting notice the output from column (2) for the model with dummy variables for the three largest FinTech market – China, United States and United Kingdom – whose positive coefficients suggest that the volume of CrediTech is much larger than what would be estimated by the considered drivers. This hints that there might be country-specific factors who intervene.

3.3 Regulation, Credit scoring and Machine Learning

The model risk management (MRM) guidelines[40] issued in 2011 by the Office of the Comptroller of the Currency, sensitive to the new innovations in computer science field, referring to the models speaks of “conceptual soundness” or “judgment exercised in model design”. At first glance it seems hindering to the introduction of machine learning methods in the industry practice, because by their nature these models relieve analysts of some manual processes, following patterns in the data that are not always evident. Of course on the other side banks and firms which desire to implement ML algorithms should open these black boxes providing more insights on what mechanisms intervene in the decision-making process, even if the transparency topic will remain for a while the weak point of machine learning.

[39] S. Claessens et al., *Op. cit.*, p. 38

[40] Office of the Comptroller of the Currency, *Supervisory Guidance on Model Risk Management*, 2011

In this regard, explainability is increasingly becoming an active area of research, different approaches that try to reverse engineer have been developed, for example[41]:

- Global surrogate model, it tries to approximate the working of the ML model through a more explainable one such as a regression or a decision trees;
- Feature importance, which is a distinctive of the Random Forest, it provides the importance of each feature, estimating how the variance of the model's prediction would change due to exclusion of a feature.
- Local surrogate model, based on selected subset of the data, different models try to approximate the complex one. A popular version of this approach is the Local Interpretable Model-Agnostic Explanation (LIME) that intends to see how the prediction changes perturbing the input data in a specific subsample.
- Instance-based approach, which provides the driving factors of the prediction in case of a specific instance (observation).
- Partial Dependence Plot (PDP), that shows graphically the impact of one or two features on the target variable.
- Leave-One-Column-Out (LOCO), that consists in removing one feature from the model and recalculate the result, if its score changes significantly it means the variable has a strong influence.

A firm specialized in ML interpretability is ZestFinance, which through math and game theory has developed algorithms able to understand the logic underlying the ML models, mainly in credit score business, due to the recent cases of discrimination imputed to an AI application. Training data could be part of the problem, because if this data are affected by existing bias of course they will be transferred also to the model. A study conducted by the Berkeley University found that machine learning systems charged Latinx/African American loan applicants higher interest rates[42].

[41] P. Bracke et al., *Machine learning explainability in finance: an application to default risk analysis*, Bank of England, 2019

[42] R. Barlett et al., *Consumer-Lending Discrimination in the FinTech Era*, Berkeley University, 2019

3.4 FinTech or TechFin?

The word BigTech refers to companies whose primary business and driving engine is technology such as Google, Amazon, Facebook, Apple, Alibaba, Tencent, Microsoft and other giants of the tech worlds. Differently, FinTech refers to technology-enabled innovation in financial services. If FinTech companies are set up to operate primarily in the financial industry, BigTechs offers financial services as part of a wider set of activities[43]. But in the recent years, defining the BigTech only tech companies can be misleading, it is enough thinking about the Apple credit card I discussed before. Bigtech companies are changing the game's rules, they are investing more and more in new markets with high grind content where technology can create value through automation. Diversification strategies have been refined, they moved from the scale economy logic to scope economy where the client is at the center and must be "monopolized". In order to do this the tech companies structurally modified the boundaries between sectors, making them increasingly fluid.

Moreover, currently the banking system is investing in R&D a fraction of the amount spent by one of the BigTech companies, in fact an analysis conducted by Supernovae Labs on the annual investments in AI reveals that the first 7 tech companies for capitalization have an overall investment capacity higher by about 50% compared to the top 500 banks by size[44].

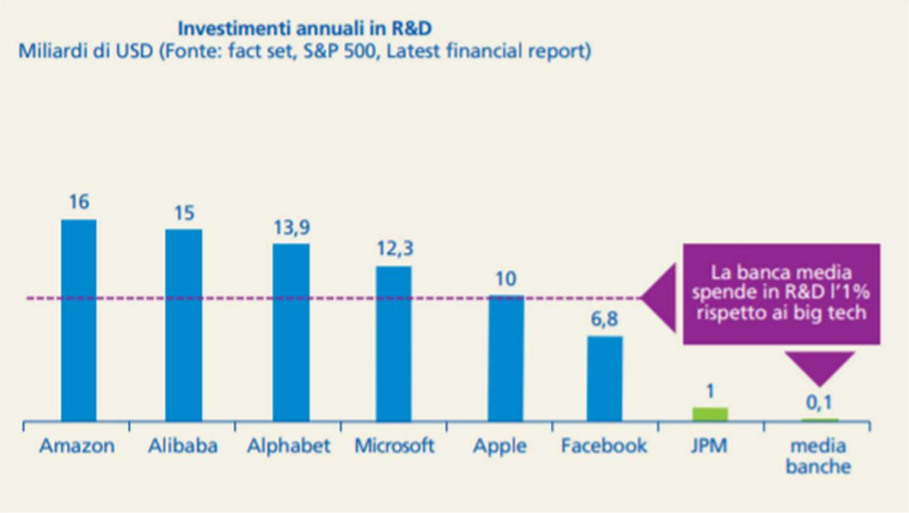


Figure 3.7 R&D investments[45]

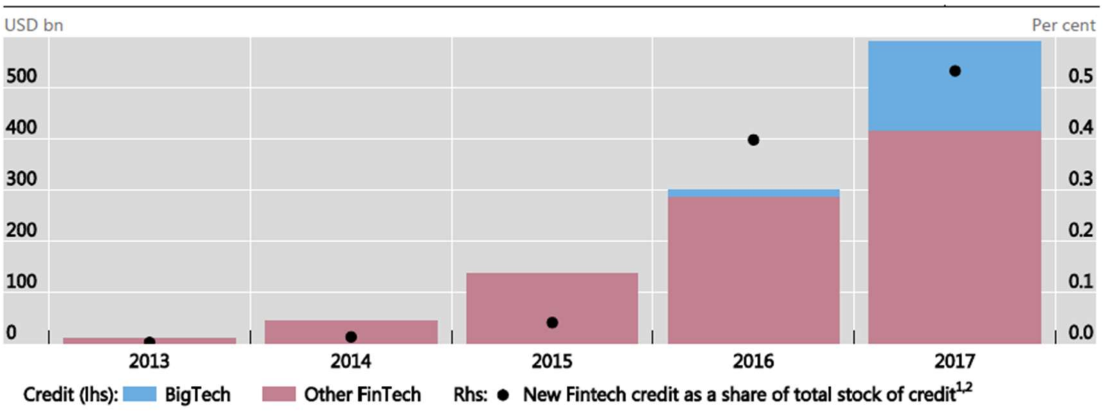
[43] BIS, *BigTech in Finance: opportunities and risks*, 2019
[44] C. Giugovaz, *Banche e BigTech: "scontro tra titani"*, Supernovae Labs, 2018
[45] Idem, p. 3

But if on one side the BigTech companies can take advantage of a large customer base and considerable constant cash flows to be directed towards investment in R&D, on the other side they have weaknesses such as the limited capacity of human relation with customers and an increasing distrust linked to the privacy topic. Banks have to increase the value added of the relation with the customer, also using AI, in order to improve the effectiveness and make an assisted service model sustainable. Traditional banking industry should consider FinTech firms as potential allies instead of a competitors, banks could compete with the tech giants merging digital innovation with solid foundations, otherwise the gap will become unrecoverable.

The BigTech firms' entry into financial services has already happened, recently they expanded into lending. It is worth mentioning the paper by J. Frost et al[46] that try to understand the drivers of BigTech credit through econometric analysis. They obtained that the existence of BigTech credit activity is positively associated with GDP *per capita*, being the latter a sort of proxy for the stage of economic development, this implies a positive correlation between BigTech activity and country's overall economic and institutional progress, but this effect become less important at higher levels of development, as observed in the FinTech case previously. Moreover the authors show how BigTech activity spreads in those jurisdictions with a less competitive banking sector, where lower costs are very attractive for consumers.

As second step Frost et al. study if the drivers of BigTech credit activity are different from those of FinTech credit, from the cross-sectional regressions emerges that the drivers are almost the same and the Fintech credit volumes are higher in countries where there is a BigTech presence, this means that the tech firms generate a fertile soil for innovations' growth but also it is interesting to note that in these countries the regulatory stringency is a more important factor. The main difference between BigTech and FinTech is that the former sees more of a boost from softer financial regulation and improved banking sector concentration than the latter.

[46] J. Frost et al., *BigTech and the changing structure of financial intermediation*, Economic Policy, 2019



The bars indicate annual global lending flows by BigTech and other FinTech firms over 2013-2017. Figure includes estimates.
[1] Total FinTech credit is defined as the sum of the flow of BigTech and other FinTech credit. This is then divided by the stock of total credit to the private non-financial sector. [2] Calculated for countries for which data were available for 2013–2017.

Figure 3.8 Global volume of new FinTech and BigTech credit[47]

The credit market is not the only concerned by this BigTech expansion, according to the estimate of Banque de France the financial services offered by the ten largest tech firms is increasing exponentially[48].

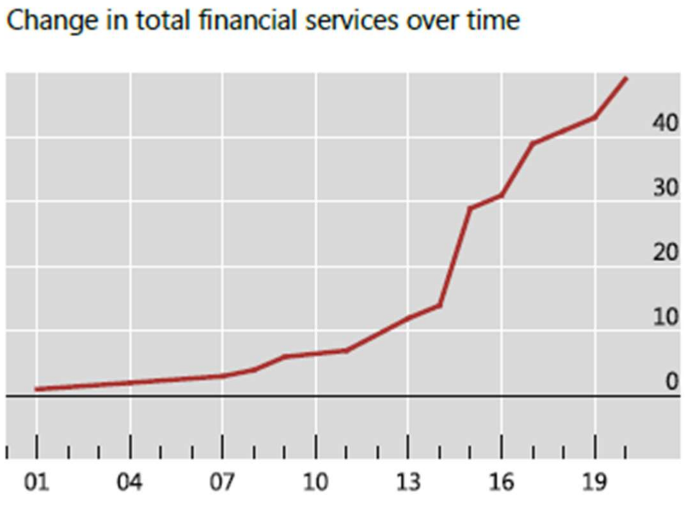


Figure 3.9 Change in total financial services over time[49]

[47] Idem, p. 24
[48] Financial Stability Board, *BigTech in Finance*, 2019
[49] Idem, p. 5

Historically payment services have been one of the first financial product offered by tech companies, to compensate for a lack of trust between merchants and customers on e-commerce platforms. This business area is particularly pronounced in China, think about Alipay which held 53.8% of Chinese market share (about $3 trillion) or WeChat. But now tech firms have diversified also in Insurance, Wealth management and Messaging services other than Credit.

Why the BigTech entered into financial services?

Even if this sector is less profitable than the tech one (considering two companies comparable for market position in both the industry, tech and financial, on the third quarter of 2019 according to MacroTrends the RoE for Apple was 53.82% while for JP Morgan was 13.21%) BigTech companies have decided to be part of it anyway.

This move can be justified by the will of:

- Diversify revenue streams.
- Access new sources of data – this kind of information can help the companies know the habits and financial positions of their clients enriching the data pool nourished by the core business.
- Complement and reinforce their core commercial activities, increasing their customer base and loyalty – they want to be perceived as all-round service.

3.5 Possible response of incumbent financial institutions

The entry into finance of BigTech will have a significantly greater impact than that of the FinTech firms for what concern competition and concentration in the financial sector. This will inevitably lead to a decrease in the margins and RoE. Banks are now called to take very important decisions for their future and survival, they can pursue different paths:

- Incumbents in financial industry can cooperate and give birth to consortia in order to share the fixed cost and gain from the combined network.
- Banks, from the smaller to the larger, can opt to partner with a BigTech (totally or relative to a particular product).
- Specialize in a niche financial service less appealing for the BigTech.

A potential optimal strategy for the banks, as stated above, could be pulling together with FinTech firms, so that the size and the customer loyalty of the bank can achieve synergies with the cutting-edge technologies of the FinTech firms. However the equation Banks + FinTechs = BigTechs is not necessarily valid, because obtaining these synergies could be a very complex process.

Other options have been experienced by banks, such as:

- redesigning internal structures in order to speed up processes and offer more timely products' developments;
- building new IT infrastructure to have a more endearing user experience;
- acquiring a "satellite" digital offer through the buyout of a FinTech to stimulate cross-selling;
- cooperating with technology companies to improve technology capabilities, getting a better data management, enhancing the efficiency of the back-office and compliance office.

Banks are facing an unprecedented phase, where paradigms are changing, structures that were believed to be immovable are now being questioned, and not all the firms in the financial industry are ready. A digital transformation is happening and it will surely leave winners and losers behind, in fact two main issues emerge: (i) IT capital expenditures sustainability in order to be competitive, and (ii) capacity to retain the customers during the digitalization.

Summarizing what has been said so far comparing benefits and risks arising from FinTech in general and BigTech in particular, from a financial stability point of view, we have:

- on one side, we have an increase in efficiency, due to cost reduction, the competitive pressure can boost the innovation and wider the access to financial services. Financial products will be cheaper, tailored and accessible.

- on the other side the increased competition can impact the profitability of the financial institutions, compromise the financial sustainability of these entities generating dangerous consequences for the real economy (Big Data and sophisticated algorithms allow BigTech to set personalized price, according to the customer's willingness to pay[50]). In case of partnership between BigTech companies and banks, operational or financial dependencies can be created, increasing the complexity of the financial system

[50] O. Barr-Gill, *Algorithmic Price Discrimination When Demand Is a Function of Both Preferences and (Mis)perceptions*, University of Chicago Law Review, 2019

and also providing new channels for risk's propagation, emphasizing the probability of a contagion in situation of operation/financial shock, also from a geographical point of view, given the global scale of these giants tech companies. Furthermore, entry into credit market by tech companies exposes us to dangers never experienced before, in fact being the BigTech subject to another regulation, we cannot know if they are able to maintain credit supply during a downturn.

3.6 Policy implications

The gateway opened by the FinTech firms in the financial industry raises a range of issues for policymakers, who are trying to adapt the regulations to this new phenomenon. Some micro-financial risks related to FinTech activities (credit, leverage, liquidity) are treated in the shadow banking policy framework issued in 2013 by FSB[51] and involving all the non-bank financial entities that are somehow close in nature to the business of traditional banking sector. Anyhow regulators are busy in designing new guidelines and standards suited for the size and the structure of domestic financial and FinTech sector. Several jurisdictions have presented regulatory test environment and innovation hubs to promote innovation but at the same time juridical changes to account for specific Fintech activities.

The financial service mainly affected by new regulations, but also the most developed, is the mobile payments. In EU for example this market has been opened to non-bank providers in 2007 through the introduction of a tailored regulatory framework specially conceived. A similar case has occurred in Japan, where in parallel deeds were issued to regulate virtual currencies and to promote cooperation between banks and FinTechs.

The protection of consumers is a main concern for policymakers, but at the same time they do not have to ignore the importance of setting up an environment that promotes the innovation growth.

[51] FSB, *Strengthening Oversight and Regulation of Shadow Banking*, 2013

3.6.1 PSD2

The PSD2 is headed in this direction, favoring competition and breaking down the banks' monopoly on the users' data. In fact the PSD2 (or Payment Services Directive 2, it follows the first one issued on 2007) is a European Union directive designed to regulate payment service and improve transparency. The new players, initially outside the scope of PSD, are now acknowledged and registered and they will access the customers' payment account (with the prior consent of the customer). This will be possible via API (Application Programming Interface) which allows the exchange of information between bank and third-party provider, that is a payment institution, introduced by the PSD2, which subject to customers' authorization offers payment services.

Among these new operators can be listed:

- Payment Initiation Service Provide (PISP), they have the opportunity, at payer's request, to activate the payment from his bank to that of the beneficiary avoiding bank's intermediation.
- Account Information Service Provider (AISP), they allow to aggregate information related to different bank accounts together.
- Card-Based Payment Instrument Issuer (CBPII), they can issue debit card associated to account of other banks and manage the money supply.

Not only APIs are thought to increase competition and innovation in financial services industry, they can be used in a more creative way, combining different kinds of service with a unique interface, in order to redefine the customer experience, making it complete[52].

The directive will also increase the security of transactions (where at least one party is located in the European Economic Area) through Strong Customer Authentication (SCA), that is planned to univocally identify and authenticate the client and the typology of transaction. It is declined in different forms:

- Knowledge – something only the user knows (password, PIN..)
- Possession – something only the user possess
- Inherence – something only the user is (fingerprint, face ID…)

[52] EFMA & Infosys Finacle, *Innovation in Retail Banking*, 2019

As well as ensuring a stronger security of transactions, PSD2 provides greater consumer rights in the event of fraud.

3.6.2 GDPR

Starting from May 25th 2018 the Regulation known as GDPR (General Data Protection Regulation) is directly applicable in every country of the EU. It derives from the challenges posed by recent technological developments and it involves the protection of people with regard to processing and transfer of personal data. This new regulation will have a significant impact on FinTech firms, in fact they are required to ask customer's consent to store and process their personal data. All the EU citizens have the permission to request that financial companies delete their personal data (*Right to be forgotten*). Furthermore, it does not matter if the headquarter is outside the EU, since the services are offered inside the EU area the GDPR compliance is demanded. Machine Learning application is affected too, the GDPR in fact restricts the use of automated decision-making, so long as it occurs without human being involved in the decision directly[53]. But the regulation specifies three cases where the automation is legal:

- The processing is necessary for contractual reasons
- It is authorized by another law
- The data subject has previously consented

3.6.3 Open banking

PSD2 introduces the idea of open banking, where data are a vital component of the migration from traditional bank to immersive and intelligent bank, which considers the customer's needs the epicenter of every decision. The CRM and the Analytics will play a fundamental role in this context, banks cannot limit to offer services requested by clients, they must be able to anticipate it, predict potential needs not yet expressed.

The key to success for banks will be to get into different aspects of a client's life, not just standing over the transactions' world but becoming a lifestyle empowerment partner. In this framework it is worth mentioning Buddybank, the mobile only bank launched by Unicredit in 2018.

[53] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016, paragraph 71

3.7 Buddybank

Buddybank, as stated before, is a digital bank and its only branch is the app that everyone who possess an iPhone can download. It is a fully interactive bank, where the relationship with the customer is managed in a personalized way through chat directly via the app. It introduces the concept of conversational banking, where the client is totally assisted 24/7 thanks to the concierge. This is the real innovation brought by Buddybank, the almost interpersonal relationship between customer and bank, that can go further merely banking issues and flow into lifestyle. This implies an inflow of transversal data about the customer that can be used to design personalized retention campaigns if the client shows an intention to close the bank account or to solidify the relationship with the customer through tailored offers.

Chapter 4

Application: Buddybank credit card default prediction with ML algorithms

During my internship in Buddybank as Data Analyst I had the opportunity to work on different projects, from the monitoring of marketing campaigns to a more sophisticated prediction model of the churn rate, that is identify the customers who could close the bank account. This is why I have decided to use Buddybank as case study for the experimental research about credit risk analysis through machine learning techniques.

At the time, the only credit product offered by the digital bank was the credit card (currently the products portfolio is wider), so after due authorization has been granted I conducted a credit risk analysis using the credit card holders dataset. Of course I know that the numerosity of the observations is not extremely high but for the purpose of the study is enough.

4.1 Dataset description

The dataset is structured and it is composed by 1711 observations of 12 variables, among which we have generally speaking qualitative and quantitative features. Referring to quantitative data I mean data that can be described using numbers and basic mathematical procedures, while qualitative data cannot be described by numbers. If we want to dig deeper we can break down the quantitative data in: continuous (data is measured) and discrete (data is counted).

So, assuming that the dataset is organized (it has a row/column structure), we now focus on what each row and column represent. Each rows represents a client that has a credit card, while each column is a variable that can describe different kinds of characteristics (demographic or banking).

4.1.1 Target Variable

The target variable is DEFAULT, and in this particular case it means the overrun at 90 days, that is an insolvency of the credit limit for 90 consecutive days, technically called delinquency. I made this choice because this event is correlated with the Default event and moreover this allows a broader sample to be analyzed given that the default as usually intended has a much smaller rate. Even with this less stringent definition of default we have very few observations of the default case class, indeed we have 144 cases (8,41% of the total). We can say the dataset is Imbalanced, but instead of using sampling methods (as Random Oversampling and Undersampling[54]) I use applications embedded in the algorithm as the *class.weight* in the Random Forest. Of course this issue will be reflected also in the selection of performance measures.

4.1.2 Features

The features, as stated before, summarize the demographical state and banking situation of each client, they try to draw a profile of the client under different perspectives. From the main variables I tried to extract insights through a phase of features transformation (for example converting quantitative variables into categorical ones, dividing all continuous numerical range into buckets) and creating artificial features, they are features not presented directly on the received data, but computed on those ones instead, also called engineered features (for example minimum number of days between subsequent deposits).

DEMOGRAPHIC	**BANKING**
ETA’	SALDO MEDIO MENSILE
GENERE (Maschio)	VARIAZIONI % SALDO
REGIONE RESIDENZA	SPESO MEDIO MENSILE
	VARIAZIONI % SPESO
	NUMERO MOVIMENTI
	NUMERO GIORNI DA ULTIMO ACCREDITO
	NUMERO MESI STESSO SALDO
	STIPENDIO
	MULTIBANCARIZZATO

[54] H. He, *Learning from Imbalanced Data*, IEEE Computer Society, 2009

4.2 Data Pre-Processing

4.2.1 Missing values

The first step of data quality is about detect and treat NAs (missing values), there can be many reasons why they occur: human errors in phase of data entry, files missing, information incomplete. Under any of these circumstances it is preferable to deal with missing values because some algorithms do not allow for them. In our case, given that the frequency of NAs is smaller than 20% (thumb's rule adopted as a practice in NAs' treatment) in the columns that present missing values, which are GENERE, ETA and NUMERO GIORNI DA ULTIMO ACCREDITO, I decided to substitute them with the mode or the median of the correspondent column. This is an approximation of course, that could lead to an under-estimation of the standard deviation[55] and also affect the relationships among variables, but in this situation we cannot afford a removal of the observations affected by missing values given the already poor numerosity of the dataset, it would entail a loss of valuable information.

4.2.2 Outliers

A similar issue is the one represented by outliers, which are even more dangerous in case of small datasets because can amplify the noise. So it is fundamental to treat them properly, in order to detect the outliers I used the following criteria: all the observations below (First quartile – 1.5 * InterQuartile Range) have been replaced by the 5th percentile, while all the observations above (Third quartile + 1.5 * InterQuartile Range) have been replaced by the 95th percentile; the rationale underlying this choice is to preserve the impact of the variable without risking to spoil the model with abnormal data.

4.2.3 Binary categorical variables

Binary categorical variables as "GENERE" or "MULTIBANCARIZZATO" have been turned into dummies.

[55] N. Mittag, *Imputations: Benefits, Risks and a method for Missing Data*, University of Chicago, 2013

4.3 Machine Learning models implementation

Once the dataset has been cleaned, is time to split it into Train and Test with a 70-30 ratio and enforcing a stratified sampling in order to keep the same proportion of the two classes : Default/No default in both sets. The idea behind this process is to train the ML algorithm with the data inside the Train set, and then employ it on the Test set and see how it performs to assess the quality of the model. This is the foundation of every predictive project in data science, in this framework we deal with supervised learning because our data are labeled and, given the available information, we are trying to predict the target variable which is known (the Default variable). Once divided the main dataset in Train and Test we can start implementing the algorithms.

4.3.1 Logistic regression

The first algorithm implemented is the Logistic regression which is a sort of generalization of the linear regression model adapted to fit classification problems. With the Logistic regression we try to predict the probabilities of class belonging given the available information set, represented by the features.

$$\pi = \Pr(\, y = 1 \mid x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + \, e^{\beta_0 + \beta_1 x}}$$

The function on the right-hand side is the logistic function and it is fundamental in this context because unlike linear regression, which assumes that y is continuous with no lower or upper ends making therefore inapplicable the concept of probability, in the logistic regression the y variable spreads on the range [0 , 1], as we can see from figure 4.1

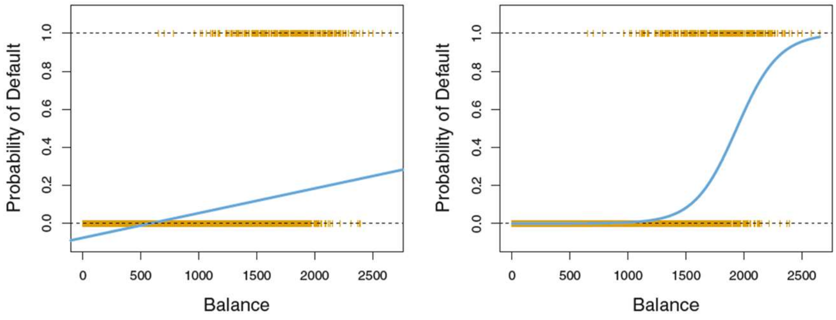


Figure 4.1 Linear regression vs Logistic regression[56]

[56] G. James et al., *Introduction to Statistical Learning*, Springer, 2013, p. 131

In order to explain in a clear way how the logistic works and what allows us to smooth the curve inside the range [0, 1] it is useful to introduce the concept of odds. The odds of an outcome occurring is the ratio of the number of ways that the outcome occurs divided by every other possible outcome instead of all possible outcomes.

To summarize:

$$Odds = \frac{P}{1-P}$$

we can see that as our probability increases, so do our odds but at a faster rate (if the probability ≈ 1 the odds will rocket to infinity). But we have a lower bound in 0, in order to get around this fact and reproduce a linear regression we have to transform the odds in log(odds), this enables us to obtain a curve that goes from minus infinity to plus infinity.

While in the linear regression β_1 represents the change in the response variable for a unit change in x, in the Logistic regression it denotes the change in the log-odds for a unit change in x.

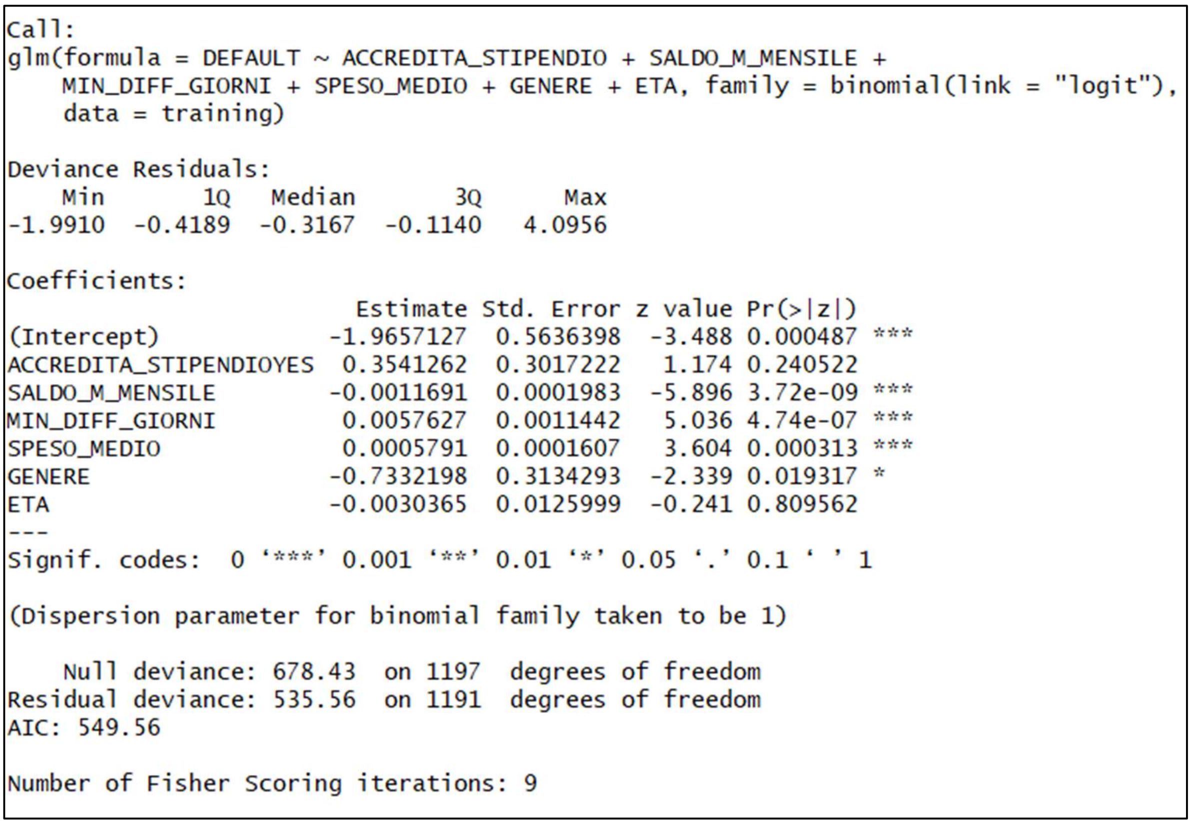
Model Results

Then I performed the Logistic regression considering only the features with a higher predictive power, detected with a Deviance analysis which consists in gradually adding each independent variable to the null model (just the intercept) and select only the features with an higher decrease rate in the Deviance[57].

$$Deviance = -2\log\left(P_{\hat{\theta}}(x_1, x_2, \dots, x_n)\right)$$

Where $P_{\hat{\theta}}$ is a probability model with estimated parameters $\hat{\theta}$ and a set of data $(x_1, x_2, \dots, x_n)$.

[57] P. Bruce, A. Bruce, *Practical Statistics for Data Science*, O'Reilly, 2017



```
Call:
glm(formula = DEFAULT ~ ACCREDITA_STIPENDIO + SALDO_M_MENSILE +
    MIN_DIFF_GIORNI + SPESO_MEDIO + GENERE + ETA, family = binomial(link = "logit"),
    data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9910  -0.4189  -0.3167  -0.1140   4.0956

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -1.9657127  0.5636398  -3.488 0.000487 ***
ACCREDITA_STIPENDIOYES  0.3541262  0.3017222   1.174 0.240522
SALDO_M_MENSILE        -0.0011691  0.0001983  -5.896 3.72e-09 ***
MIN_DIFF_GIORNI         0.0057627  0.0011442   5.036 4.74e-07 ***
SPESO_MEDIO             0.0005791  0.0001607   3.604 0.000313 ***
GENERE                 -0.7332198  0.3134293  -2.339 0.019317 *
ETA                    -0.0030365  0.0125999  -0.241 0.809562
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 678.43  on 1197  degrees of freedom
Residual deviance: 535.56  on 1191  degrees of freedom
AIC: 549.56

Number of Fisher Scoring iterations: 9
```

Figure 4.2 Logistic regression summary

Variable Importance

We can remark that:

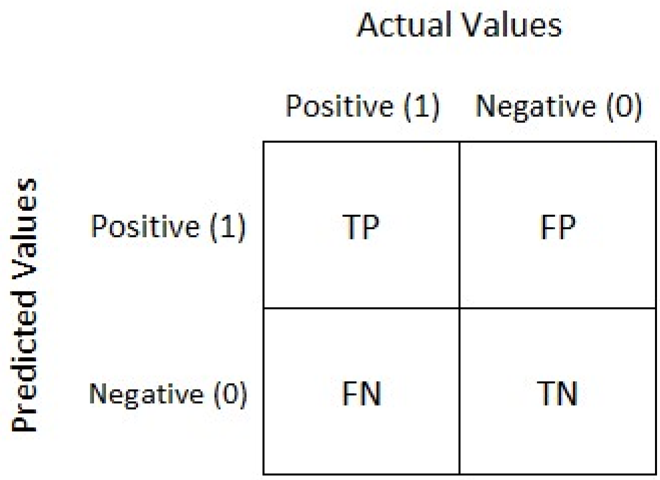
- SALDO_M_MENSILE has a negative sign coefficient, which means that ceteris paribus a client with higher account balance is less likely to default then someone with a lower account balance.
- MIN_DIFF_GIORNI suggests us that increasing the number of days from the last form of deposit in the bank account increases the likelihood of default.
- GENERE has a negative sign coefficient and this tells us that being male with all other variables kept equal decrease the likelihood of default compared to being female; but we can also notice that the significance is smaller than the other features, it might be worth investigating the reason behind this result, maybe I am missing some information related to the gender variable (like the salary, which unfortunately is an unavailable data for many clients and notoriously is lower for women).

Model evaluation

The key point, in order to evaluate the “quality” of the model, turns around testing the model on the test set, whose data were not used in the process of training. Therefore the performance measured on this test set should be considered a good indicator of how well the model will perform on future data, this practice is commonly called Hold-out Sampling. Sometimes besides train and test sets a third sample is created, the validation set, which is used to tune the hyper-parameters of the algorithm, so instead of optimizing the parameters directly on the train sample they are set using unseen data with the validation set and finally the model performance with hyper-parameter tuning is assessed on the test set. In my case this could result redundant due to the low numerosity of the dataset with a consequent high risk of overfitting.

Confusion Matrix

Once enacted the prediction on the test set, we can show the results with a Confusion Matrix, a very practical table that shows where the model does mainly wrong.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 4.3 Confusion Matrix

This minimal table is very useful because it contains all the information about the most frequent kind of errors made by the model: Type I error (false positive), the client is predicted to default while he/she did not, Type II error (false negative), the client is predicted to not default while he/she did. In order to obtain the Confusion matrix it is necessary to set a threshold, a cut-off point in the predicted probabilities above which we have a positive result and negative otherwise. Setting this threshold manually involves discretionally increasing/reducing the FP and FN, that sometimes can be hoped and preferable but usually not.

ROC Curve and AUC

In order to not arbitrarily decide the threshold we build the ROC (receiver operating characteristic) curve, which is a graph showing the performance of a classification model at all the thresholds. This curve plots two parameters:

- True Positive Rate (Recall)

$$TPR = \frac{TP}{TP + FN}$$

- False Positive Rate

$$FPR = \frac{FP}{FP + TN}$$

The ROC graph depicts relative tradeoffs between benefits, true positives, and costs, false positive. The point (0,0) indicates the strategy of never predict a positive classification, while the point (1,1) represents the opposite strategy, unconditional issuing of positive classification. The perfect classification is constituted by the point (0,1)

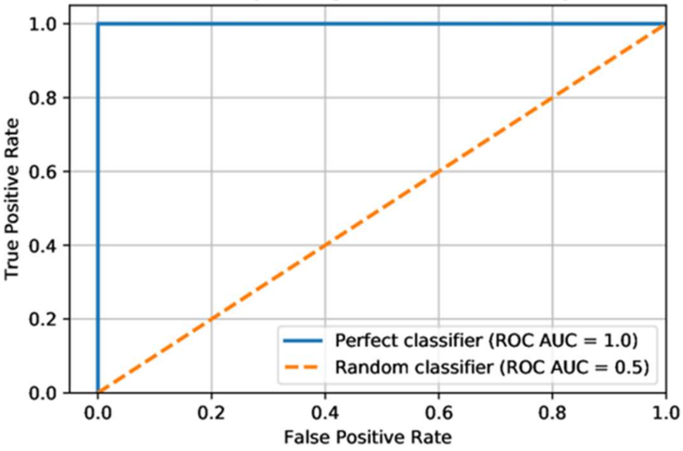


Figure 4.4 ROC Curve

Classifiers that appear on the south-left side of the graph are usually considered “conservative”, they need strong evidence to classify an instance positive, thus they make few false positive errors but also have a low positive rate. The diagonal dotted line represents the strategy of randomly guessing a class, that is the model cannot get information from the data. It is used as basic benchmark to assess the predictions. Intuitively the more the ROC curve pushes on the top-left of the chart the better it is.

A very useful property of the ROC curve showed by T. Fawcett in his paper *An introduction to ROC analysis*[58], is its insensitivity to changes in class distribution. In fact if the proportion of positive to negative observations changes in a test set, the ROC curve does not change, while other performance metrics such as accuracy, precision and F score would be affected.

Changes in class distribution are not unrealistic, for example in fraud detection, proportions of fraud varied significantly over time[59] .

To measure the performance across the all possible thresholds we need to compute the Area Under the Curve (AUC) which measures the ability of the model to distinctly separate the positive cases from the negative ones[60].

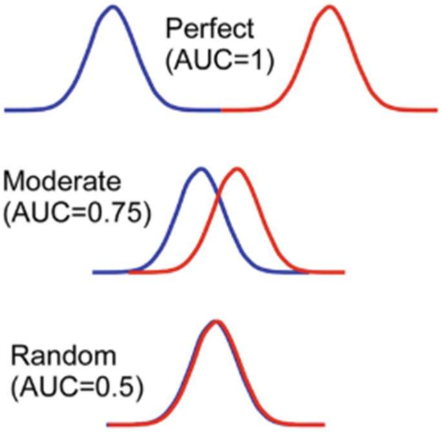


Figure 4.5 AUC cases

[58] T. Fawcett, *An introduction to ROC analysis*, Pattern Recognition Letters, 2006
[59] T. Fawcett, F. Provost, *Adaptive fraud detection*, Data Mining and Knowledge Discovery, 1997
[60] M. Sokolova, G. Lapalme, *A systematic analysis of performance measures for classification tasks*, Information Processing and Management, 2009

Model performance valuation

The ROC plot and the corresponding AUC for the Logistic regression are the following:

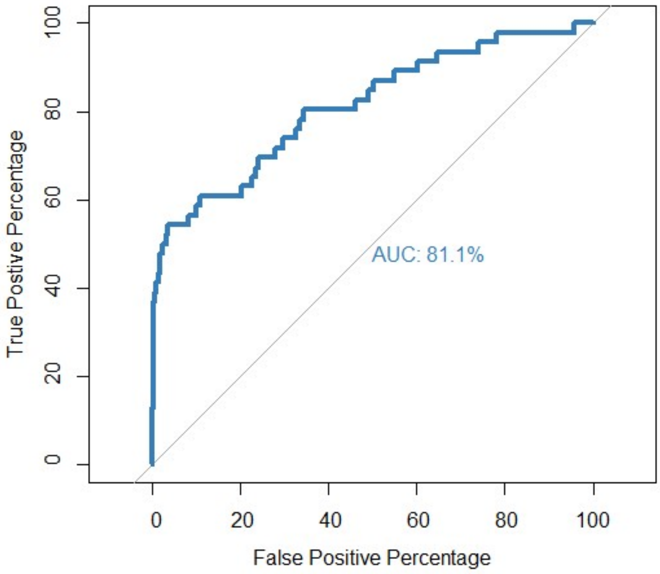


Figure 4.6 Logistic regression ROC and AUC

In this way it is not necessary to arbitrarily fix a threshold to evaluate a model, because all the possible threshold are tried to measure the performance of a model.

4.3.2 Random forest

The second algorithm performed is the Random Forest, I used the Ranger package in R which is a fast implementation of random forest that reduces significantly the computation time. But before explaining how this algorithm works it is necessary to describe what is a decision tree algorithm that is the pillar of the Random Forest.

Decision Trees

Tree-based methods are very simple and easy to understand, however they are not competitive with the best supervised learning approaches. Decision trees can be applied to regression and classification problems, in this framework the focus is on classification. Decision trees are built using a heuristic called recursive partitioning, which is also commonly known as *divide and conquer* because it exploits the available features to split the data in smaller and smaller subsets of similar classes. The mechanism is very simple, at the beginning where we consider the entire dataset the algorithm selects the feature that is most predictive of the target class. Then based on distinct values of this feature the instances are divided in different groups and the algorithm keeps doing this node after node, choosing gradually the best candidate feature, until a stopping criterion is reached:

- Almost all the observations at each node are part of the same class
- There are no more features left
- The predefined size limit of the tree has been reached

In their paper A.E. Khandani et al[61] apply a Classification And Regression Tree model to analyze credit risk using customer transactions and credit bureau data from January 2005 to April 2009 for a sample of a major commercial bank's customers. They propose a traditional measure of consumer credit risk that combines credit factors such debt-to-income ratios with consumer banking transactions. In order to address a common issue in credit-default data, the unbalanced proportion of bad and good realizations, and to improve the predictive power of the model, they use a technique called "boosting". This algorithm, also called AdaBoost, iterates several decision trees, base learners, and each time recomputes the weights of the observations, especially the missed one. The weight for observation *i* at the *n*th iteration is given by:

$$w_i^{(n)} = w_i^{(n-1)} e^{[\alpha_{n-1} I(f_{n-1}(x_i) \neq y_i)]}$$

where $I(\cdot)$ is an indicator function that indicates whether the model has correctly predicted the outcome y_i given the input vector x_i, $\varepsilon_{n-1}$is the weighted average error of the model from the *(n-1)*th iteration while the re-weighting coefficient α_{n-1} is defined as:

$$\alpha_{n-1} \equiv \ln \left(\frac{1 - \varepsilon_{n-1}}{\varepsilon_{n-1}}\right)$$

[61] A.E. Khandani et al., *Op. cit.*

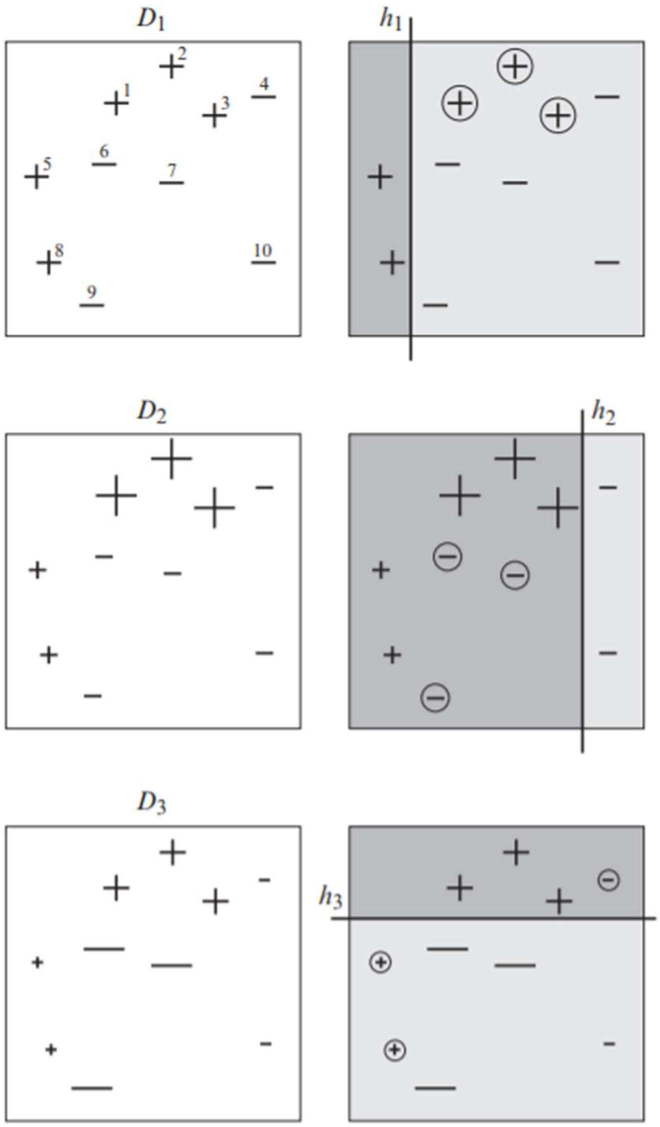


Figure 4.7 AdaBoost algorithm[62]

As we can notice from Figure 4.7, at every iteration *i* the base learner makes an hypothesis h_i based on the data distribution D_i in order to identify the positive observations, that is the darker area. All the misclassifications, the circled observations, are re-weighted in the following iteration *i+1*, resulting in a greater impact on the hypothesis h_{i+1}.

This is why the algorithm, that helps to improve the performance, is called AdaBoost, which stay for "adaptive boosting".

[62] R.E. Schapire, Y. Freund, *Boosting: Foundations and Algorithms,* The MIT Press, 2012

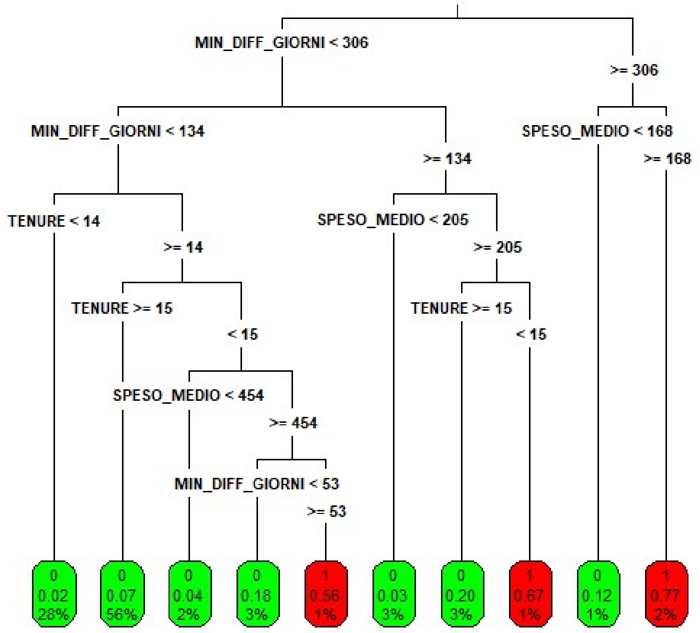


Figure 4.8 Decision Tree with Buddybank sample

Figure 4.7 shows an example of a decision tree using the *rpart()* package[63] and the same features of the Logistic regression. As we can see at the root node the feature selected by the algorithm is MIN_DIFF_GIORNI (due to its importance), after this node the dataset is split in two subsets, one where the values for the MIN_DIFF_GIORNI variable are higher or equal than 306, and another one that does not respect the first rule, so we gradually move on to the next nodes and we add more rules until getting to the final predictions. We can think of the decision as a series of if-else statements that step by step make the whole dataset more granular.

The stopping criterion is very important otherwise the tree keeps growing with too specific splitting rules (as we can see in Figure 4.7) and increasing exponentially the risk of over-fitting, for this reason we talk about tree pruning; the parameters we can set are the *minsplit* (number of observations to appear in a leaf before it is pruned) and *maxdepth* (maximum number of nodes in a branch).

Anyway, we can notice that the decision tree is very explainable, which is a strength of this algorithm, in fact we can derive all the decision rules applied to the dataset and understand the underlying rationale. The terminal node tell us: the predicted class, the Gini index for that particular subset and its percentage of the all dataset.

[63] T. Therneau, B. Atkinson, *Recursive Partitioning and Regression Trees*, CRAN, 2019

Gini Index

The Gini index is defined by:

$$G_m = \sum_{k=1}^{K} \widehat{p_{mk}}(1 - \widehat{p_{mk}})$$

Assume we are trying to compute the Gini index for the feature GENERE when the region *m* is male, $\widehat{p_{mk}}$ represents the proportion of training observations in the *m* region that are from the *k*th class (default, no default). Combining the results from every region, this metric measures the total variance across the *k* classes for a given feature, more extreme is the value assumed by $\widehat{p_{mk}}$ (near 0 or 1) stronger will be the partition between the classes for that region and lower will be the Gini impurity index. At every node the algorithm iteratively selects the variable with the lower Gini index, that is with the better ability to discern.

So, why we do not use the Decision Tree algorithm?

Because they do not have a strong level of predictive accuracy, as stated before, this method can perform well in the training set but poorly with new data given the inability to effectively generalize a classification rule. In fact decision tree is highly exposed to the bias-variance tradeoff, if the tree grows too much overfitting is inevitable (high variance), on the other side if we try to avoid it pruning the tree the accuracy can be affected (high bias).

The solution to this dilemma is Random Forest, which is an ensemble technique aimed to reduce variance without increasing bias, using multiple decision trees. The basic concept behind Random Forest is the "wisdom of crowds", in fact a considerable number of relatively uncorrelated trees will outperform any individual decision tree model. This is possible by training on different samples of the data and by using a random subset of features that insure the low correlation among the single models, allowing the trees to protect each other from individual errors[64].

Describing in detail these two key points:

1. Bagging (or Bootstrap aggregation) – given a set of n independent observations with variance σ^2, the variance of the mean is given by $\frac{\sigma^2}{n}$, so averaging a set of observation reduces variance. We can apply this concept simply building separate prediction models using different training sets and averaging the resulting

[64] L. Breiman, *Random Forests*, 2001

predictions. Each training set is created through bootstrap, by taking repeated samples (with replacement) from the whole dataset. This makes possible obtaining *N* different bootstrapped training sets. We just have to train our model on the *N* sets and finally average the predictions:

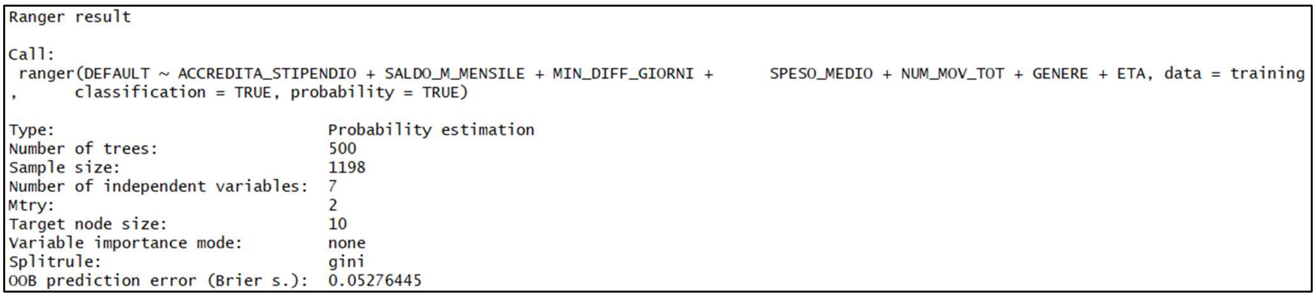
$$f_{bag}(x) = \frac{1}{N}\sum_{n=1}^{N} f^{n}\,(x)$$

In this case we have *N* different decision trees, each of which predict a class (0 or 1) and the final step is take a majority vote: the overall prediction is the most commonly occurring class among the *N* predictions.

2. Random feature selection – while in the decision tree in every node the algorithm considered every possible feature and pick the one with the lower Gini index (so the more predictive one) in the random forest each tree can select only from a random subset of features (usually the square root of the variables' number) and then apply the Gini impurity criterion. This results in a lower correlation across trees, a more diversification and a better performance.

Model Results

Performing the Random Forest in our analysis we obtain the following results:



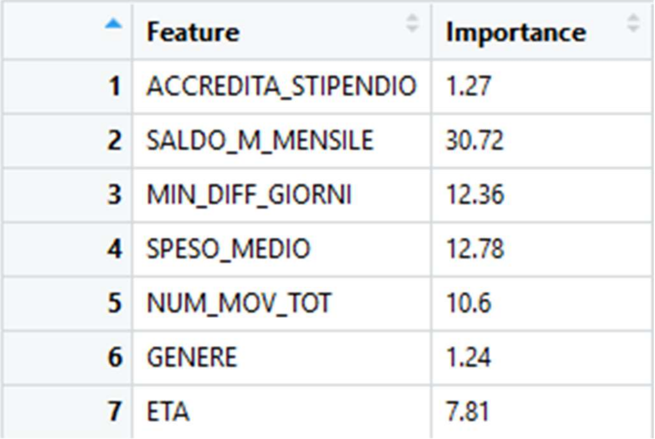
```
Ranger result

Call:
 ranger(DEFAULT ~ ACCREDITA_STIPENDIO + SALDO_M_MENSILE + MIN_DIFF_GIORNI +      SPESO_MEDIO + NUM_MOV_TOT + GENERE + ETA, data = training
,      classification = TRUE, probability = TRUE)

Type:                             Probability estimation
Number of trees:                  500
Sample size:                      1198
Number of independent variables:  7
Mtry:                             2
Target node size:                 10
Variable importance mode:         none
Splitrule:                        gini
OOB prediction error (Brier s.):  0.05276445
```

Figure 4.9 Random Forest summary

As we can notice from the model output, I have set to 500 the number of trees that compose the forest, and left the default setting for the number of features randomly selected for each tree (the square root of the number of independent variables). Moreover we have a sort of test error, the OOB (Out-of-Bag) error, which gives a preliminary valuation of how the model performs with new data.

Variable Importance

Studying the variables importance (computed using the mean decrease in Gini Index and expressed relative to the maximum) we can notice that the results in the Random Forest model are similar to the ones related to Logistic Regression, SALDO_M_MENSILE is the most predictive feature.



	Feature	Importance
1	ACCREDITA_STIPENDIO	1.27
2	SALDO_M_MENSILE	30.72
3	MIN_DIFF_GIORNI	12.36
4	SPESO_MEDIO	12.78
5	NUM_MOV_TOT	10.6
6	GENERE	1.24
7	ETA	7.81

Figure 4.10 Random Forest variable importance

Model performance valuation

If we try to predict the observations in the test set in order to measure the quality of the model we obtain the following ROC and AUC:

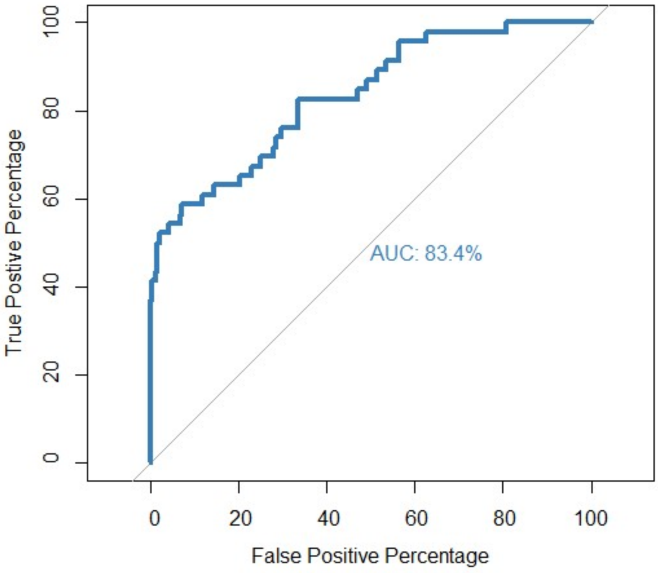


Figure 4.11 Random Forest ROC and AUC

4.3.3 Neural Networks

The last algorithm that I perform in this analysis is the Neural Networks, which is also the most complex and as the name suggests it imitates the human brain. The main idea behind this algorithm is that through a single (or multi) layer networks of neurons we can basically approximate every function and this is very useful when we do machine learning because we need to estimate the function that links our features with the target variable. The layers are composed of nodes, which combines input from data with associated coefficients (called weights) that either magnify or dampen that input. Then these input-weight products are summed in the net input function and subsequently passed through the so-called activation function, to assess whether and to what extent that signal should move on through the network to affect the final output. Figure 4.11 describes how a Neural Networks looks like.

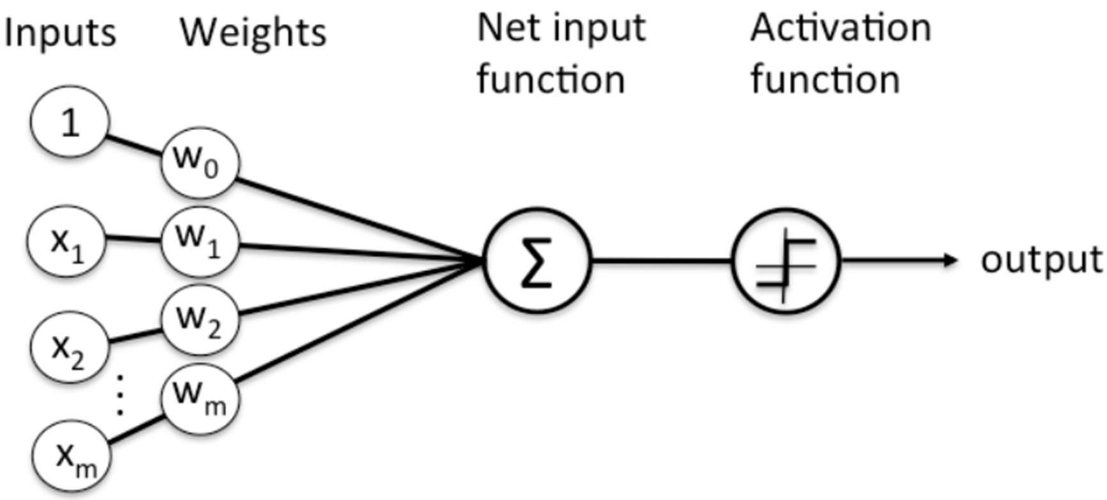


Figure 4.12 Neural Networks diagram

A node layer is a series of these neuron-like switches and each layer's output is also the subsequent layer's input.

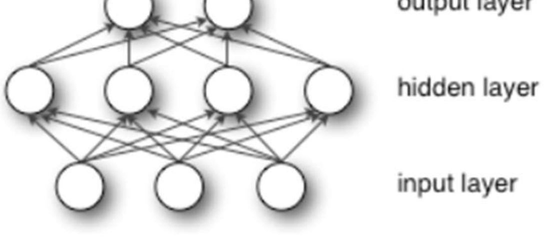


Figure 4.13 Neural Networks Layers' structure

In order to explain in the most understandable way this concept I think could be very effective focus on an extremely simple neural network. Suppose we have only one feature and one hidden layer made of one node.

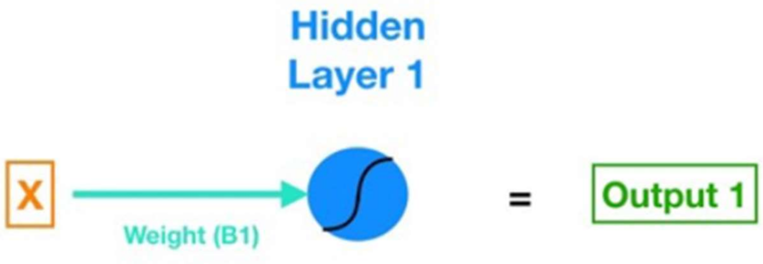


Figure 4.14 Neural Networks with 1 node Hidden Layer

To express the previous diagraph in mathematical notation we can say:

$$Sigmoid(B_1 * X + B_0) = Output\ 1$$

Examining each element we have:

a. X is our input, the only feature we have
b. B_1 is the estimated parameter (weight)
c. B_0 is the Bias, every neuron has its own bias
d. $Sigmoid$ is our activation function (a sigmoid function has a monotonic S-shaped curve, examples are the logistic regression or the hyperbolic tangent, and we use it in order to restrict the results between 0 and 1)

Activation function

The activation function allows us to get the output of the node, it converts the inputs in the results. We use the Logistic function because the purpose of this research is to predict the default probability, so we have a range from 0 to 1, and this function fits it perfectly.

This entire process is called forward propagation and in a normal neural network is enacted multiple times (depending on the number of neurons in each hidden layer).

Given this structure the aim of the algorithm is to minimize the cost function (which represents how wrong our predictions are respect to the target outcome), that is finding the set of weights and biases that minimize it. In order to do this we enforce the gradient descendent, an iterative optimization algorithm used to minimize any function moving in the direction of steepest descendent, determined by the negative of the gradient. The gradient of a function is the vector whose elements are its partial derivatives with respect to each parameter, that tell us how the cost function would change if we applied a small change to any weight or bias (our parameters).

Backpropagation

Now we have all we need to introduce the backpropagation, which is the fundamental step in neural networks. After the forward propagation, in order to optimize our prediction we want to compute the error attributable to each neuron, starting from the layer closest to the output all the way back. The size of the error of a particular neuron with respect to the other neurons' error is proportional to the impact of that specific neuron's output on the cost function, that is why we use the partial derivatives of the cost function with respect to the neuron's parameters. The rationale behind this is that if a neuron has a more considerable error than the others then resetting the weight and the bias of that specific neuron should have a bigger impact on the model's total error. In poor terms we can say that the backpropagation enables us to calculate the weighted error for each neuron and to compute the partial derivatives (the gradient vector) so we can exploit the gradient descent.

Cost function

In order to measure the performance of a ML model we need a Cost function, which quantifies the discrepancy between predicted and actual value. So the Cost function is what we want to minimize with the gradient descent. In the binary classification framework the Cross-Entropy function is the practice, even if sometimes depending on the application of the deep learning other losses are preferable, as in case of highly noised input data the squared hinge loss would be more suitable[65].

[65] K. Janocha, W.M. Czarnecki, *On Loss Functions for Deep Neural Networks in Classification*, Cornell University, 2017

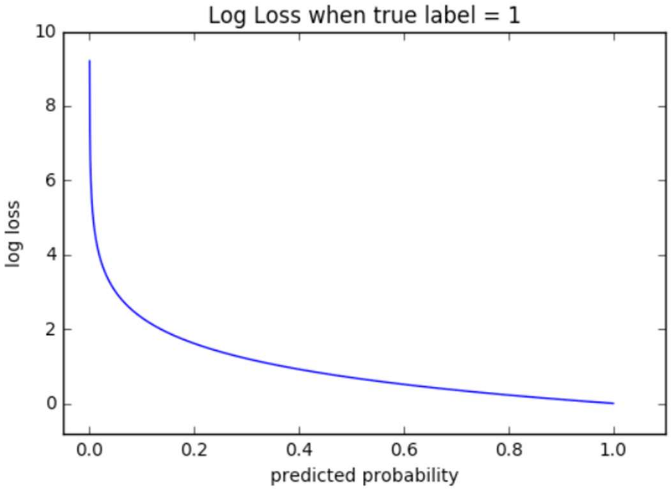


Figure 4.15 Cross-Entropy Loss function

The Loss function is then given by:

$$C = -\sum_i p_i \log q_i = -p_i \log q_i - (1 - p_i) \log(1 - q_i)$$

where p_i is the actual label, zero or one, and q_i the predicted probability [0,1].

The more the predicted probability diverges from the actual label, the higher is the Cross-Entropy and, through the properties of this cost function[66], the faster the neurons will learn. This can be proved doing the partial derivative of the Cross-Entropy function with respect to the weight, consider the previous example of the one feature one node hidden layer neural networks, given that:

$$C = -p_i \log(\frac{1}{1 + e^{-(wx+b)}}) - (1 - p_i) \log(1 - \frac{1}{1 + e^{-(wx+b)}})$$

Where $\frac{1}{1+e^{-(wx+b)}}$ is what I previously called q_i, we obtain:

$$\frac{\partial C}{\partial w} = \sum x(\frac{1}{1 + e^{-(wx+b)}} - p_i)$$

and we can easily note how the error in the prediction controls the learning rate.

[66] M. Nielsen, *Neural Networks and Deep Learning*, 2019

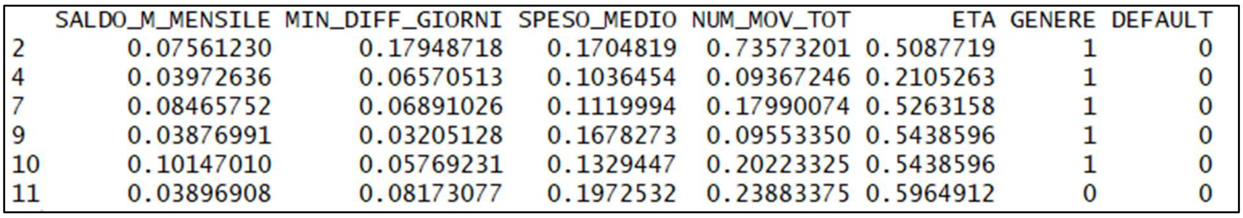
Data Normalization

While constructing the neural network model one of the most important procedure is: Data Normalization, adjusting the values measured on different scales to a common scale. So I proceeded to transform the data, without outliers, using a min-max normalization, which does not affect the dummy variables:

$$\frac{x - \min(x)}{\max(\mathrm{x}) - \min\ (x)}$$

So our data now are capped between 0 and 1 as it emerges from Figure 1.9 that shows the first 6 rows of the normalized train set:

- head(nn_train)



```
   SALDO_M_MENSILE MIN_DIFF_GIORNI SPESO_MEDIO NUM_MOV_TOT       ETA GENERE DEFAULT
2       0.07561230      0.17948718   0.1704819  0.73573201 0.5087719      1       0
4       0.03972636      0.06570513   0.1036454  0.09367246 0.2105263      1       0
7       0.08465752      0.06891026   0.1119994  0.17990074 0.5263158      1       0
9       0.03876991      0.03205128   0.1678273  0.09553350 0.5438596      1       0
10      0.10147010      0.05769231   0.1329447  0.20223325 0.5438596      1       0
11      0.03896908      0.08173077   0.1972532  0.23883375 0.5964912      0       0
```

Figure 4.16 Normalized train set

Model Results

I set one hidden layer and the parameter *linear.output*[67] equal false, otherwise we would obtain a linear regression model, in fact this setting enables the activation function to work. Figure 4.16 shows the plot of the Neural Networks obtained.

[67] S. Fritsch, F. Guenther, *Training of Neural Networks*, CRAN, 2019

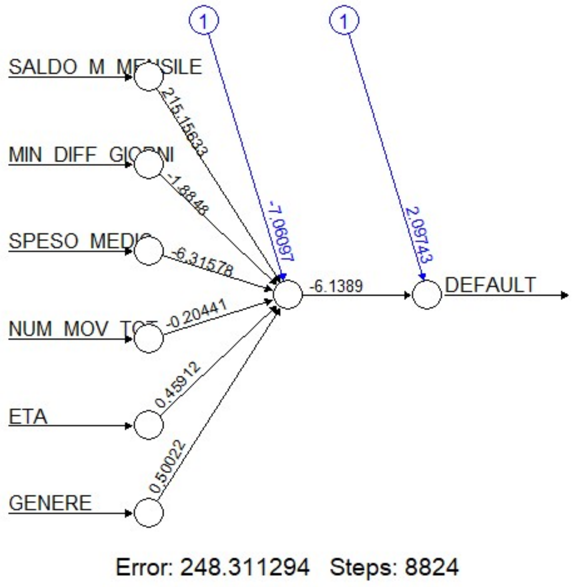


Figure 4.17 Neural Networks summary

From the previous plot we can notice that to each feature and bias (the blue node) has been attributed a weight which will contribute to the estimate of the final prediction. The steps represent the number of iterations occurred to train the neural networks over the entire training set.

Model performance valuation

The AUC and the ROC curve for the Neural Network model are the following:

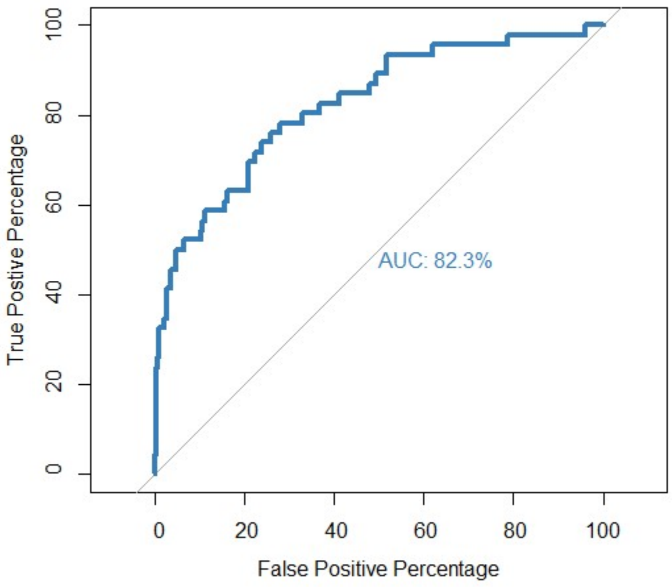


Figure 4.18 Neural Networks ROC and AUC

4.4 Comparing the models

In order to summarize all the results showed in this section and compare the models it can be useful to plot the predictions' accuracy in terms of AUC, in a single ROC graph, so that we can realize which is the best one.

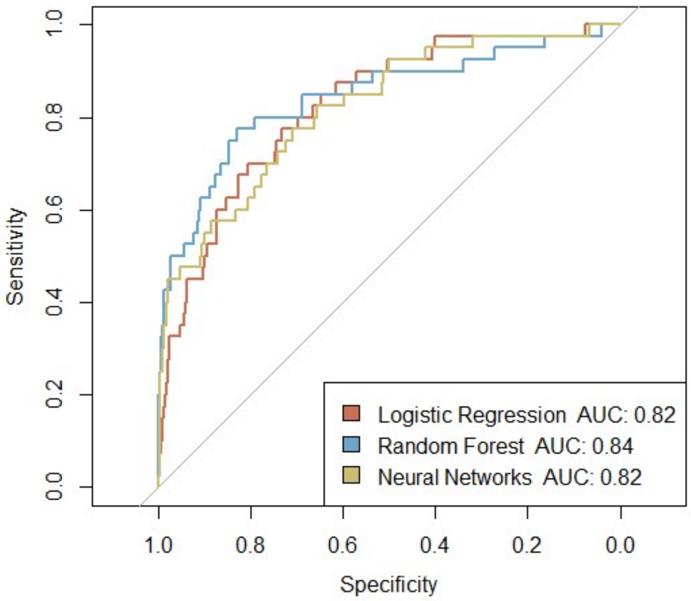


Figure 4.19 Comparison models' ROC and AUC

In this version of the ROC we compare Sensitivity (True Positive Rate) with Specificity (True Negative Rate) to describe the performances of the different models and it is easy to see how the Random Forest model slightly outperformed the Logistic Regression and Neural Networks. The study conducted by S. Lessmann et al. obtains the same results of the analysis, showing how advanced methods can outperform simple classifiers in credit scoring, in fact the Random Forest algorithm obtained an higher score than Logistic regression, Linear discriminant analysis and also Neural Networks[68].

Of course this result does not imply that Random Forest algorithm is generally better than Neural Networks, on the contrary the latter being very complex is usually more effective. In literature, the term "complexity" is used to describe the possible outputs a model can generate in relation to the possible inputs.

[68] S. Lessmann et al., *Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research*, European Journal Of Operational Research, 2015

The choice of the algorithm depends on the data you have and on what is your goal, in this context with few data an ensemble technique as Random Forest can result more robust than a simple Neural Networks (increasing the number of hidden layers was exaggerated given the dataset's dimensions). The criteria that lead the algorithm's decision are Performance and Robustness[69], in terms of performance the Neural Networks has the higher potential, but attached with it there is also the risk of excessive adaptation to the training sample. So when assessing the performances, reliability plays a decisive role from the decision-maker point of view, but also the comprehensibility is a key factor, especially in a field as credit risk where the regulation imposes full transparency.

The sophistication of Neural Networks (NN) algorithm is reflected also in the model implementation phase, in fact during the pre-processing NN requires that missing value are filled, categorical variables are converted into numerical and features are scaled otherwise larger values are treated as more important in the training. Moreover in the training step, different parameters need to be set in order to have a proper model, it is necessary to define the Neural Network structure, that is how many layers to use and then how many neurons in each layer, the activation function, and lastly the training algorithm and its related parameters. On the contrary, Random Forest has a very straightforward implementation that, considering its performance and robustness, makes it very appealing. Obviously with a large dataset it can be valuable commit time in the hyper-parameter tuning, which can improve significantly the model performance.

4.5 Extension of the model

A possible extension of the model could include the use of alternative data, being Buddybank a digital bank with a mobile app I could add features related to the "digital behavior" of the customer (frequency of app login, interaction with the concierge, tone of the conversations through sentiment analysis). Moreover the lifestyle function can enrich my data pool with information inherent to the personal life of customers, which could be a strong predictor.

Lastly, with the introduction of PSD2 Buddybank can get access to data related to customers' bank account with other institutions, improving the knowledge about its clients' financial

[69] J.D. Kelleher et al., *Fundamentals of Machine Learning for Predictive Data Analytics*, The MIT Press, 2015

situation. In this regard, the study conducted by Jagtiani and Lemieux[70] confirms the usefulness of exploiting alternative data inside decision making processes or machine learning models, keeping in mind that some set of alternative data may work very well for sample of consumers but may be not representative for others.

I have not undertaken this path because of the low numbers of observations, but in future with a larger dataset it could be very interesting analyze how these variables impact the outcome and see if they really have any predictive power.

[70] J. Jagtiani, C. Lemieux, *Op. cit.*

Conclusion

The main goal of this thesis is to exhibit the potential of Artificial Intelligence in Finance, describing current trends in FinTech industry and applying a real case-study.

Reports from several financial institutions, such Financial Stability Board and European Banking Authority, outline the increasingly important role played by FinTech firms in the existing financial ecosystem and therefore the necessity for a more detailed regulation of this phenomenon, considering that estimates from EBA indicate the 31% of European FinTech firms as not subject to any authorization or registration regime. According to the annual Global Analysis of Investment in FinTech by KPMG released on February 2019, the global investment activity in 2018 (considering Venture Capital, Private Equity and Merger & Acquisition) has been around $120 billion, with China, United States and Europe as drivers of the growth. This implies that the FinTech revolution is expected to grow and traditional banks need to promptly react in order to avoid a gradual reduction of their market share.

In such framework Machine Learning has a pivotal role because the quantity of data generated and treated on a daily basis is exponentially increasing. Analytics is essential in processing data and mining useful information to design a tailored and appealing product in businesses such payments services or lending. Focusing on the retail credit area, the experimental research conducted in this thesis tries to address the following task: compare the predictive power of credit scoring models based on modern Machine Learning techniques with that of a more traditional approach, in particular correctly classify the credit card Default customers. Implementing a Logistic Regression, an ensemble model such as Random Forest and a Neural Network on the same Buddybank dataset, emerges that the Random Forest method outperforms the others. It has a positive discrepancy (with respect to the other models) of two percentage points in terms of AUC, implying a considerable better ability to rank a randomly chosen positive instance higher than a randomly chosen negative instance. It can effectively identify to which class unseen observations belong although the low numerosity of the training sample. One possible reason of this result can be attributed to the properties of the Random Forest:

- bootstrap aggregation
- random feature selection

that make this method particularly robust to scarcity of data and to overfitting.

Accordingly, Machine Learning is an extremely valuable tool that financial institutions are working to employ it more consistently across different areas, as shown by the estimates of Bank of England in its report *Machine Learning in UK financial services* (2019). At the same time a proper regulation is needed in order to exploit the real capacity of ML and to promote a fair use of it.

Recommendations for future research

The next step in Machine Learning application to credit scoring would be the use of alternative data which do not refer to traditional banking information but rather to non-credit information as insurance payments or utilities payments. The recent introduction of the PSD2 offers the opportunity to get access to different sources of alternative data and therefore to have a more complete picture of the financial situation of a loan applicant. A proper combination of the two types of information would enormously benefit the banks on one side and allow a financial inclusion of those previously excluded because considered too risky on the other side.

Bibliography

Altman, E. I. et al., 1981. *Application of Classification Techniques in Business, Banking and Finance*, JAI Press

Altman, E. I. and Saunders, A., 1997. *Credit risk measurement: Developments over the last 20 years*, Journal of Banking & Finance

Atiya, F. A., 2001. *Bankruptcy Prediction for Credit Risk Using Neural Networks: A Survey and New Results*, IEEE Transactions on Neural Networks

Bank of England, 2019. *Machine learning in UK financial services*, Financial Conduct Authority

Barlett, R. et al., 2019. *Consumer-Lending Discrimination in the FinTech Era*, Berkeley University

Barr-Gill, O., 2019. *Algorithmic Price Discrimination When Demand Is a Function of Both Preferences and (Mis)perceptions*, University of Chicago Law Review

Basel Committee on Banking Supervision, 2005. *An Explanatory Note on the Basel II IRB Risk Weight Functions*

Beatty, A. and Liao, S., 2011. *Do delays in expected loss recognition affect banks' willingness to lend?*, Journal of Accounting and Economics

Bank for Iinternational Settlements (BIS), 2019. *BigTech in Finance: opportunities and risks*

Bracke, P. et al., 2019. *Machine learning explainability in finance: an application to default risk analysis*, Bank of England

Breiman, L., 2001. *Random Forests*, Machine Learning

Bruce, P. and Bruce, A., 2017. *Practical Statistics for Data Science*, O'Reilly

Claessens, S. et al., 2018. *Fintech credit markets around the world: size, drivers and policy issues*, BIS

Cohen, B. H. and Edwards Jr, G. A., 2017. *The new era of expected credit loss provisioning*, BIS

European Financial Management Association, 2019. *Innovation in Retail Banking*

Eisenbeis, R. A., 1977. *Pitfalls in the Application of Discriminant Analysis in Business, Finance, and Economics*, Journal of Finance

European Commission, 2018. *Ethics Guidelines for Trustworthy AI*

European Parliament, 2018. *General Data Protection Regulation*, paragraph 71

Ewanchuk, L. and Frei, C., 2019. *Recent Regulation in Credit Risk Management: A Statistical Framework*, MDPI

Fawcett, T., 2006. *An introduction to ROC analysis*, Pattern Recognition Letters

Fawcett, T. and Provost, F., 1997. *Adaptive fraud detection, Data Mining and Knowledge Discovery*

Financial Stability Board, 2019. *BigTech in Finance*

Financial Stability Board, 2013. *Strengthening Oversight and Regulation of Shadow Banking*

Fisher, R. A., 1936. *The Use of Multiple Measurements in Taxonomic Problems*, Annals of Eugenics

Fritsch, S. and Guenther, F., 2019. *Training of Neural Networks*, CRAN

Frost, J. et al., 2019. *BigTech and the changing structure of financial intermediation*, Economic Policy

Gambacorta, L. et al., 2019. *How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm*, BIS

Giugovaz, C., 2018. *Banche e BigTech: "scontro tra titani"*, Supernovae Labs

He, H., 2009. *Learning from Imbalanced Data*, IEEE Computer Society

Jagtiani, J. and Lemieux, C., 2019. *The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform*, Wiley

Jagtiani, J. and Lemieux, C., 2018. *Do FinTech Lenders Penetrate Areas That Are Underserved by Traditional Banks?, Journal of Economics and Business*

James, G. et al., 2013. *Introduction to Statistical Learning*, Springer

Janocha, K. and Czarnecki, W. M., 2017. *On Loss Functions for Deep Neural Networks in Classification*, Cornell University

Kelleher, J. D. et al., 2015. *Fundamentals of Machine Learning for Predictive Data Analytics*, The MIT Press

Khandani, A. E. et al., 2010. *Consumer credit-risk models via machine-learning algorithms*, Journal of Banking & Finance

Lapalme, G. and Sokolova, M., 2009. *A systematic analysis of performance measures for classification tasks*, Information Processing and Management

Lerner, A. P., 1934. *The Concept of Monopoly and the Measurement of Monopoly Power*, The Review of Economic Studies

Lessmann, S. et al., 2015. *Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research*, European Journal Of Operational Research

Lu, H., 2013. *Credit Scoring Model Hybridizing Artificial Intelligence with Logistic Regression*, Journal of Networks

Luger, G. F. and Stubblefield, W. A., 1997. *Artificial Intelligence: structures and strategies for complex problem solving*, Addison Wesley Publishing Company

Mitchell, T. M., 1980. *The Need for Biases in Learning Generalizations*

Mittag, N., 2013. *Imputations: Benefits, Risks and a method for Missing Data*, University of Chicago

Navaretti, G. B. et al., 2018. *FinTech and Banks: Friends or Foes?*, European Economy – Banks, Regulation, and the Real Sector

Ng, A., 2018. *Machine Learning Yearning*

Nielsen, M., 2019. *Neural Networks and Deep Learning*

Odom, M. D. and Sharda, R., 1990. *A Neural Network Model for Bankruptcy Prediction*, International Joint Conference on Neural Networks

Office of the Comptroller of the Currency, 2011. *Supervisory Guidance on Model Risk Management*

Ozdemir, S., 2017. *Principles of Data Science*, Packt

Resti, A. and Sironi, A, 2007. *Risk Management and Shareholders' Value in Banking*, Wiley

Schapire, R. E. and Freund, Y., 2012. *Boosting: Foundations and Algorithms*, The MIT Press

Therneau, T. and Atkinson, B., 2019. *Recursive Partitioning and Regression Trees*, CRAN

Tse, T. C. M. et al., 2019. *The AI Republic*, Lioncrest Publishing

Turing, A. M., 1950. *Computing Machinery and Intelligence*, Mind

Wolpert, D. H. and Macready, W. G., 1997. *No Free Lunch Theorems for Optimization*, IEEE Transactions on Evolutionary Computation