Bocconi University

Bachelor of Science in Economic and Social Sciences

Academic year 2019/2020

LENDING SMART:
THE POTENTIAL OF DATA ANALYTICS IN ASSESSING CREDIT RISK

by

Enrico Guffanti

Thesis advisor: Professor Stefano Caselli

Milan (Italy), July 2020

This work is dedicated:
To my father, solid reference point and long-time comparison.
To my mother, for all her care and advice.
To my brother, lifelong adventure teammate.
To my whole family, those who have been there since day one.

To myself.

Special acknowledgements go to Professor Stefano Caselli
for being my advisor on this thesis work.

TABLE OF CONTENTS

INTRODUCTION

Data. Everything seems to be revolving around data nowadays, to the point that this word may already be earning the title of "golden word" of the current decade, if not also of the next one. But while the accolade can be understood to having been entrusted by popularity, the comparison to gold sounds extremely accurate, certainly more on the basis of their intrinsic value rather than their scarcity.

But why is data so valuable? Essentially, data is an asset that many companies already own, and sometimes in very large amounts, therefore it is not unthinkable to believe it could be put to some good use, not only to draw extremely detailed insights on business matters, but also to generate an enormous competitive advantage.

With this work, we will focus on credit risk and on how lending changed after the Great Recession of 2008, when fear and uncertainty about repayments by borrowers became an even more major concern. At that time, the role of the rising industry of Analytics and its tools for Big Data handling and management turned out to be crucial in partly mitigating those concerns, while at the same time opening up to unprecedented opportunities for many companies to completely re-think their whole business models, from simple day-to-day activities to as much as their whole organizational structure. Across this work, specific attention will be devoted to the industry of peer-to-peer lending and the ground-shaking impact that Credit Risk Analytics has brought: to do so, we will consider the case of LendingClub, and eventually take on the challenge of building from scratch an optimized model for credit risk assessment and prediction in a CRISP-DM framework, by using data by the same company. Prior to all this, however, an introduction to the value of Business Analytics and its relevance for financial intermediaries will be given, along with a presentation of the advantages it offers in the context of lending.

THEORY SECTION

What is credit risk? (1) To start off the theory section of this work, we shall begin from introducing the very central topic of credit risk. Formally, credit risk is the possibility of a loss that a lender incurs, resulting from the fact that a borrower may be failing to repay a loan wholly or meet other contractual obligations. Not repaying means that the borrower is not returning what he owes to the lender according to the contractual agreements, which in general involve principal and interests. Credit risk comes up every time lending is involved, both at the private and corporate level, some examples being mortgages offered to households, credit cards, credit given by a company to a customer, bond issuers making payments upon request, or insurance companies paying for a claim. In each of these cases, the lender may not (and shall not) assume the probability of getting repaid (every time or by each borrower) according to the scheduled terms to be 1, and this is the very practical manifestation of credit risk. As any other kind of risk in the business world, the goal is to provide the most accurate measurement of credit risk prior to the loan being granted, and eventually denying it in case of extremely high uncertainty of a repayment. Such calculations are then based on the borrower's overall ability to repay a loan according to its original terms, an assessment that involves looking at the borrower's credit history, capacity to repay, capital, the loan's conditions, and associated collateral. If there is a higher level of perceived credit risk, investors and lenders usually demand a higher rate of interest for their capital.

Credit risk vs. Default risk (2) A specification for the above, however, is needed, as by looking at the definition of credit risk we immediately grasp how it refers to a situation in which the borrower does not respect any of the contractual terms agreed with the lender regarding the repayment to be made. This is very different from solely stating that it refers to the risk that the borrower will never be repaying the lender, as such money could still come back, but later than the time scheduled upon signing of the contract. This apparently small detail in the interpretation of a definition is actually fundamental, both conceptually and for the analyses that will be performed later on in the context of this work: the distinction between credit risk and default risk.

Default risk is defined as the possibility that the borrower will be unable to make the required payments on his debt obligation. While risk of default exists in virtually all forms of credit extensions, it is now clear how it differs from the aforementioned credit risk, as here we are considering the fact that lenders and investors should look at the probability of never receiving a repayment for their loan as being different from 0. In this sense, default risk is a special case of credit risk in which the contractual breach by the borrower comes in the form of impossibility of repaying the loan at any time, present or future.

As with credit risk, also default risk assessment gets performed prior to the loan approval by the lender, and is measured in probabilistic terms, according to which higher default risk leads to a higher required return, and in turn, a higher interest rate.

In light of what is said above, a specification on the analyses that will follow is important, as the Predictive Model presented in the empirical part of this work will not be considering default risk: due to sampling specifications and data proportions that will be fully explained in the pages to come, taking on the challenge of predicting defaulting borrowers against compliant ones would make the model highly inaccurate, mainly due to the fact that results would be obtained on an extremely small sample size, while the models that will be adopted for predictions require bigger datasets to be trained on. For this reason, records that refer to defaulting borrowers in the dataset at hand have been excluded from the sample for the predictive analysis, consequently eliminating the assessment of default risk from this setting. Therefore, as an initial note, the analyses that will be performed later on try (among other things) to best predict credit risk, leaving out default risk.

Lending after the Great Recession of 2008 As stated in the introduction of this work, Analytics started becoming more and more popular among companies on the sunset of the first decade of 2000s, the same time that will forever be remembered as an infamous one in history due to the financial crisis that hit on a worldwide scale. Following the Great Recession in 2008, the financial industry came under greater pressure to improve risk systems and models to reduce future losses and recurring crises, with many financial institutions realizing it especially with regard to credit risk, where conventional methods for its management had proven insufficient. Many financial players back then were looking for more advanced and innovative approaches to manage credit risk, and data analytics was certainly one of the most powerful resources available, and just at the right time. But while these new tools were getting widespread and bringing back hope to a much troubled industry, it represented also a disruptive event that would have changed radically the way of doing business and approaching corporate challenges. In times of fast technological innovation, the revolution by many financial institutions was to place their customers at the heart of their digital transformation, quickly and efficiently capturing a vast amount of information on those same borrowers in order to estimate the credit risk associated to each and every one of them. Contrary to what had been done in the past, banks and other financial players were not only considering general and easily accessible customer data, such as demographics and information on current and previous credit lines, but they were now going steps further: from spending patterns to late/defaulting payments, passing from evaluating less traditional data sources, such as activity on social

network platforms and interactions with physical branches or call centers. In this new era, minimizing credit risk means that financial intermediaries have to arm themselves with the power of being able to obtain a 360° view of their customers' financial behavior. Put it like this, it all may sound quite intimidating to any normal client applying for a loan and mindful about his own right to privacy, but a deeper pondering of the implications of profiling sheds positive light on some advantages for customers as well: examples are the increased chances for those customers without credit history of getting a loan approved [(3)], or a more seamless and rapid credit granting with reduced turnaround time.

Data and Analytics At this point, however, before delving deeper into how analytics techniques can matter in the context of credit risk scoring, it is necessary to take a step back, a detour allowing to define some key points that stand as foundations for the whole work.

First of all, regardless of the context taken into consideration, data is always a source of insights. Obviously raw data by themselves most unlikely deliver meaningful findings, and this is because their value needs to be uncovered and extrapolated by means of (more or less) advanced techniques in the context of a process called "Information Value Chain". This process takes raw data, more or less structured, as inputs, extracts information, which then gets converted into knowledge: now data is "useful", in the sense that it carries practical meaning and can be used to make decisions and take actions, which reflect into profit. Possessing data then turns out to be a necessary, but not sufficient, condition to increase profits.

Data is certainly something not scarce in today's world. Data floods from any possible side and source for those who want to collect it and have ways of storing it. The incredible advancements in modern technology offer a number of opportunities for capturing such an immense amount of data, such that referring to this time in history as "Big Data Era" feels more appropriate than ever. Big Data are characterized by the 5-Vs:

1. Volume, as the attribute "Big" is not a mere commercial term to draw attention from the media, as here we are really considering data at scale, an enormous and always growing one indeed, in the order of terabytes or petabytes per data warehouse.
2. Variety, as data may come in many forms, some examples being structured, semi-structured, unstructured, text, multimedia, etc. …
3. Velocity, since we are dealing with data in motion, which often undergoes a real time analysis by means of powerful tools.

4. Veracity, a term addressing the uncertainty that comes with any data types, and whose management of their reliability and predictability of their inherent imprecision are core topics for many companies.
5. Valuable, the core characteristic that justifies the collection of data. Information is the real "diamond in the rough" to be mined out of massive databases.

Now that we described what "Big Data" are, we also need to introduce the science that is adopted to handle them in the context of corporations with the goal of using them to increase profits: Business Analytics. As Thomas Davenport and Jeanne Harris have defined it in one of the most influential works for this industry, Business Analytics is about "the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions"[(4)]. Analytics on Big Data has been shown to solve a number of business problems ranging across many fields, from customer churn prediction to social network analysis, from recommendation engines to employee performance analysis, passing from weather forecasting for business planning and sentiment analysis. More closely related to the scopes of this work, however, are the major applications of Business Analytics in finance, coming in the form of pricing analysis, fraud detection, personalized banking, stock price forecasting, and advertising analysis.

Also, there exist multiple types of Analytics, depending on the specific flavor that the analysis may take. Among them, we mainly recall:

1. Descriptive Analytics: This type of Analytics looks at historical data to highlight useful and insightful characteristics and associations between variables, along with patterns in the data. In this sense, the attention is focused on a "rear view" over the information extracted from the dataset at hand. Among the most important techniques falling under this category, we can cite Clustering Analysis and Association Rule Mining.
2. Predictive Analytics: This kind of Analytics, instead, takes the opposite approach with respect to the first one right above, as here the direction of extrapolated insights is a "forward view". By studying the existing patterns in the current dataset, the adopted model takes care of predicting the outcome of one or more target variables whenever a new, never-seen dataset is analyzed by the same model. Depending on the type of dependent variable included in the dataset, Predictive Analytics takes the form of either Classification (if the outcome to be forecasted is categorical) or Regression (if the target variable is continuous).

3. Prescriptive Analytics: This kind of Analytics is the most comprehensive one, as some kind of projects may require both Descriptive and Predictive insights over the data.

By looking at the above categories and what the goals are whenever each of them gets implemented, one may argue that all the glitter surrounding the new science of Analytics may be just boiling down to a re-interpretation of something that has been carried out forever in the context of the corporate world. While this is partly true, modern Business Analytics is something that goes far beyond traditional Business Intelligence, and not only because nowadays companies can leverage the power of Big Data. For example, traditional Business Intelligence may be overlapping almost perfectly with Descriptive Analytics due to their common rear view orientation and the use of methods that mainly include reporting through KPIs and metrics. Also, they both deal mainly with structured data, and the knowledge generation process is manual: as a consequence, business initiatives undertaken with the support of traditional Business Intelligence are reactive. Modern Business Analytics, instead, while still incorporating insights on the past, also has a future orientation, mainly due to the implementation of methods like predictive modeling, data mining, simulation, and optimization. Additionally, it mainly deals with both structured and unstructured data types; and most importantly the knowledge generation process is automatic: therefore business initiatives in modern Business Analytics terms are proactive.

Other points of distinction between traditional techniques and modern ones lie in the production and consumption of analytics, and in the role of the analyst. In the past, organizations have employed analysts in specific roles to define and explore business models, analyze data, and produce reports and dashboards. Analysts were the main producers of analytics, providing insights into the organization's performance and supporting the decision-making processes of the management, who was then consuming analytics to make decisions and drive actions. Today, however, with the rise of increasingly sophisticated self-service analytic capabilities and the flattening of organizational hierarchies, the production of analytic insights is supported by a growing spectrum of people across (and also outside) the organization: producers are increasingly more operational workers, entrusted with decisional power within their own areas of responsibility, in addition to executives and managers. This very fact highlights a shift in the corporate dynamics brought forward by the advent of modern Analytics, also resulting in the lines between producers and consumers of analytics becoming increasingly more blurred, marking the birth of "prosumers" of analytics.

The value of Business Analytics [(5)] By now, the idea that data carry value with them is widely understood to be reality, as information can be an asset that has had and will continue to have

high relevance in the business world. But while being smarter has almost always meant being successful, success requires more than just knowledge of statistical techniques or ways of dealing with Big Data. Most importantly, a fundamental condition for companies to benefit from implementing analytics in their own business model is an understanding of how all this translates into competitive advantage. What may sound as an automatic process, it is in fact not, with a number of pitfalls that must necessarily be anticipated to maximize the probability of success, mostly relevant:

1. Companies often struggle to effectively communicate the core insights hidden inside all the data they have amassed, and this is not a minor concern. The questions that companies should constantly be asking themselves are: "Are the methods and techniques in place the best ones for our case? Is the value creation procedure adopted wringing every last drop of value from all our business processes?". Working on data is not enough to maximize success of analytics initiatives if the tools that are being applied (obviously without additional injection of new funds) leave room for further improvements: in this sense, analytics does translate into competitive advantage, but only for those who can fully exploit its potential.
2. Difficulties in even quantifying a priori both the value of business analytics initiatives themselves and the value generated by those same projects is another issue, as the calculation of costs and benefits is heavily dependent on factors changing over time, or even having an aleatory component. All this has the potential of being able to considerably influence the success rate of analytics initiatives, just like (if not more than) any other corporate project.
3. Efficient and effective delivery of findings is one more problem whose importance should not be minimized, and for a very key reason: how can complex and convolute metrics of success rate shed light on how competitive advantage is being created? The art of working with Analytics is to be able to use the most complicated and advanced modeling techniques, while at the same time selling their value in the most straightforward and appealing way.

Competing on Analytics [(4)] A natural question that should come up at this stage is: "Then why should I decide to invest in Analytics if it is not even possible to perfectly quantify its value a priori?". An answer that requires no effort would be of stating that "it is what everyone else seems to be doing, so why not do the same?", but is it really the case? If from the point of view of rational agents working their way through a setting of uncertainty this is the safest equilibrium, incorporating more variables allows for some very deep and real-world considerations on challenges that corporations need to face, but which may fall unnoticed after having described all the shiny advantages that Analytics brings about.

Deciding to let Analytics through the doors of a company requires much more than a standards Prisoner's Dilemma approach to be properly addressed, and while playing it "the conservative way" and failing to innovate may result in advantaging competitors in the long run, "the innovative way" could prove hampering over the course of a sour short run.

[(6)]As any corporate initiative, also Analytics implementation requires funds, and not a few. Even if database architectures and software are increasingly becoming more affordable, the investments required to put in place everything that is needed may reach far in the order of several millions. Not only that, as Analytics initiatives are definitely not a plug-and-play technology: as said, if a company wants to capitalize on its innovations, it has to be willing to wait a while before being able to see conspicuous returns, and for a number of reasons. First of all, companies must be aware that hardware is not enough: data infrastructure, AI software tools, vast processing power, and data storage are just worth a piece of equipment without data expertise, modeling development abilities, and possible support from outside the organization. Most importantly, we are in a time where most businesses are not born digital, and this is an additional challenge that highlights how company's way of working, structure, and especially culture need to be aligned and reshaped to welcome the adoption of Analytics with minimum resistance. This last point represents the "devil in the detail" of this situation, as impediments to innovation may well come from inside the organization, from those same employees that, called to update their technological skills and accept working side-by-side AI, may be fearing of losing their jobs in favor of the latter.

Another major concern that companies for sure need to address is the fact that those millions of funds that have to be channeled towards analytics could well be allocated on some other corporate initiatives, possibly bringing returns sooner and carrying less uncertainty with regards to their value. A crucial factor to be considered here is prioritization of corporate objectives, and where innovation in technology (with all its related consequences) ranks on the list of all possible investments.

Moreover, considering that the majority of AI transformations take between 18 and 36 months to complete, with some extending to as late as 5 years before their full achievement, the matter of opportunity-cost in light of the aforementioned investments is, again, challenged. To be noticed is also the fact that this time estimates can hardly be shortened in practice, as databases take time to be filled up, and the data collection step of the whole process is inevitably time-consuming.

[(4)]Finally, since databases were mentioned, it is not to be forgotten how competing on analytics means competing on technology, which implies that competitors need to closely monitor and push the IT frontier, which mainly boils down to:

1. Computing hardware needing to be constantly kept up-to-date, as the volumes of data required for analytics applications may strain the capacity of low-end computers and servers.
2. Putting in place a data strategy, since data has to be stored, integrated, presented in standard formats, and made accessible through proprietary databases to those who need it within the organization.
3. Investing in the latest cybersecurity available to protect the value of own data.

This section of the work for sure went in the opposite direction compared to all the other ones, challenging the idea that implementing Analytics is always beneficial. But while for sure it was highlighted how the reality of these kinds of projects is not something that abstracts from corporate dynamics and constraints, both financial and cultural, the idea that those who succeed in completing their organization's transition to analytics have gained a solid competitive advantage came out even stronger. It is now clear how companies that excel at implementing AI will find themselves at a great advantage in a world where humans and machines working together outperform either humans or machines working on their own.

The impact of Analytics on financial intermediaries: Models that matter Over the course of the previous pages, we have introduced the topics of data and Business Analytics referring to a general framework, thus the considerations that were made could fit on many different types of organization. Now, however, it is interesting to apply those insights to the finance industry, especially by inspecting the advantages for financial intermediaries from using Analytics. Many uses have been implemented in the financial sector, mainly to support the creation of competitive advantage in terms of optimization of operations and support to strategic decisions, and with the principal aims of increasing revenues and growth, improving efficiencies, lowering costs, and allocating investments. Not only that, because analytics proved to be a powerful arm to fight off and differentiating from competitors on grounds other than quality, boosting customers' satisfaction and maximizing customer retention, thus enhancing company's reputation. Not to be forgotten is how analytics allows for a better understanding of the drivers of financial performance, along with an assessment of the impact of non-financial factors and a closer monitoring of risks (reflecting in lowered costs due to excessive uncertainty, and quicker detection of frauds). Related to this last point and to the topic of this work, Predictive Analytics can turn out to be

very useful when it comes to credit risk scoring applications, where the algorithm would be called in to calculate how big of a risk the financial intermediary would take in case it chose to underwrite each specific new customer asking to borrow. In addition to that, Descriptive Analytics would help to identify those customers having the greatest profit source potential, so that particular products or offerings could be tailored and proposed to them. This process of learning from historical data could be extremely sophisticated in determining customers' reservation prices and highest willingness to pay, which in the context of modern personalized banking and lending could be key for drafting the most efficient contract possible, perfectly balancing customers' incentives and needs.

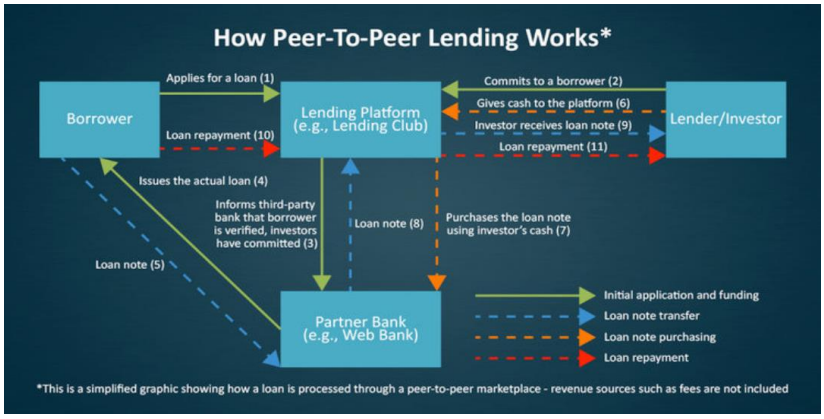
But successful implementations of Analytics do not stop here, as financial intermediaries came up with further ways of leveraging the power of data and modeling, for example through:

- Simulation of multiple alternatives/scenarios and determination of unexpected constraints;
- Forecasting of future trends in the financial markets to help customers buy/sell stocks and make more informed decisions regarding their financial portfolios;
- Constantly monitoring of the compliance to all the necessary regulations, mainly involving the limitations of risks, for a faster and more accurate response in case of troubles;
- Enhancement of everyday activities, such as (for banks) the perfect allocation of cash at each ATM location.

The peer-to-peer lending industry and Data Analytics [(7)] Entering the core of this work means narrowing the spectrum of analysis and focusing on addressing a more specific, yet extremely insightful, setting. As a matter of fact, the topic of credit risk and that of Data Analytics find the perfect blending in the context of the peer-to-peer lending industry, which will be the framework of interest that will be taken into consideration throughout the remaining sections.

Peer-to-peer lending refers to the practice of lending money to individuals or businesses through online services that match lenders with borrowers: such platforms then mainly become tools that get together people or companies with complementary interests. On the one side, lenders and investors provide their funds by injecting them into the lending platform; while on the other side borrowers in need of financing can apply for a loan by entering the network of that same lending platform. Peer-to-peer marketplaces achieve their scopes by usually teaming up with a partner bank and make profits by essentially collecting a brokerage fee for every time they successfully connect investors and borrowers, thus they are incentivized to increase the total number of transactions taking place on their platforms. But where is

Analytics in all this? This is actually a crucial point for peer-to-peer lending companies, as advantages of implementing analytics are extremely sound, on both shores of the money flow: while on the funding side lenders can select the best opportunities available for investing their funds and earning higher returns as compared to savings or traditional products offered by banks; borrowers can instead obtain loans at exceptionally convenient interest rates. Not only that, as the whole process devotes special attention to keeping credit risk under control, mainly through the continuous and unceasing effort of AI techniques in assessing which borrowers are eligible for appearing on the list of possible recipients of a loan, and consequently be selected by investors.



The case of "LendingClub" [(8)] In the context of peer-to-peer lending companies, a specific one will extensively be considered throughout the rest of this work: not only because it represents an incredible story of success in the industry that is being considered, but also since it portrays a remarkable example of Analytics being used to enhance a business model. This company is "LendingClub Corporation" (https://www.lendingclub.com/).

LendingClub is an American peer-to-peer lending company, headquartered in San Francisco, California. Founded in 2006 as one of the very first firms in the industry, it is now the world's largest peer-to-peer lending platform, after having transformed the way people access credit by bringing borrowers and investors together, helping millions of people take control of their debt, grow their small businesses, and invest for the future. The company went public on December 10th, 2014 (NYSE: LC; https://www.nyse.com/quote/XNYS:LC), raising $ 1 billion in what became the largest technology IPO of 2014 in the United States. LendingClub also partners with WebBank, a FDIC-insured state-chartered industrial bank headquartered in Salt Lake City, Utah for support on underwriting operations.

An exceptional role model indeed, as LendingClub claimed it originated up to $ 15.98 billion in loans as of December 2015, and generated no less than $ 574.5 million in revenues over the course of 2017, while accounting for as much as 1,837 employees still in the same year. On top of this, the company also became the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market.

In the spirit of a FinTech company, also at LendingClub the corporate culture is centered on the belief that accessing credit should be seamless, a target achieved through committing to making borrowing and investing simple and easy, keeping costs low and opportunities high. In this sense, investors can browse the loan listings on LendingClub's website and select those loans that they want to invest in based on the information supplied about the borrower, who are being assessed real-time by Analytics tools. As a consequence, investors make money from optimized investments, borrowers save money on interests, and LendingClub makes money by charging borrowers an origination fee and investors a service fee.

[(9)]Among those recognizing LendingClub as an amazing opportunity for borrowing and investing smarter, a name that does not end up unnoticed is that of Google LLC. In fact, also Google has teamed up with LendingClub to invest some of its massive cash holdings into the small-business borrowers market, looking at this partnership as a better opportunity as compared to attempting building the necessary architecture to try and reach the same objective on its own, thus possibly becoming a FinTech competitor itself. The choice by companies like Google can also be justified in light of viewing investments in loans through Lending Club as a way to grow own capital faster than, for example, money market funds, while at the same time reinvesting in product partners.

Customers' satisfaction at the center While by this point the advantages of using Analytics in the context of lending appear as evident as they could be, many financial players (especially traditional banks) tend to underestimate the role of data: the value-extraction process is deemed important, yet at the same time is often not prioritized. It is true how, in fact, in the scenario of digital innovation and disruption of the traditional financial industry, other steps of the process may look more compelling to the management's eyes, for example the creation of online portals, setting up of apps for mobile interactions, or instituting policies to ensure regulatory compliance when moving lending processes online. Without disregarding the importance of such initiatives, developing a highly performing risk analytics software proves fundamental, both in light of allowing a necessary level of automation and also for creating a positive customer experience. As a matter of fact, customer expectations have changed with the rise of digital services, and banks cannot afford to follow slow, arduous and manual processes to make every loan decision. Not only that, but what would be the advantage from a customers' point of view to apply for a loan online if they still need to have their request go through a long trail of workers making decisions? A system that automatically identifies relevant data pertaining to a loan application, and quickly determines if conditions make the customer eligible for it, is actually a critical element of the whole digital transition process and its successful aftermath.

EMPIRICAL SECTION

After having introduced the theoretical framework of the work, it is time to take on a more empirical approach in order to demonstrate that those concepts proposed in the previous sections are not just something that happens in books, but are in fact extremely related to real-world dynamics in the industry of Business Analytics. The value of Big Data is something that exists, and its possible impact on corporate activities will be presented in the remaining pages of this work.

The idea from now on is to introduce the reader to the CRISP-DM analytics framework and to guide him throughout each and every step of it. This general scheme for facing Business Analytics problems is something that has been implemented for a long time by many successful companies: to demonstrate what actually goes on in practice in the corporate context, the presentation of the framework will be accompanied by the construction of an analytics model on a dataset from LendingClub. All the analyses will be performed using RapidMiner Studio, a data science software that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. In RapidMiner, the analyses on (Big) data are performed by creating a tailored analysis process (.xml or .rmp), which is a sequence of building blocks connected by means of links. Each of those building blocks is called "operator", and has the following general characteristics:

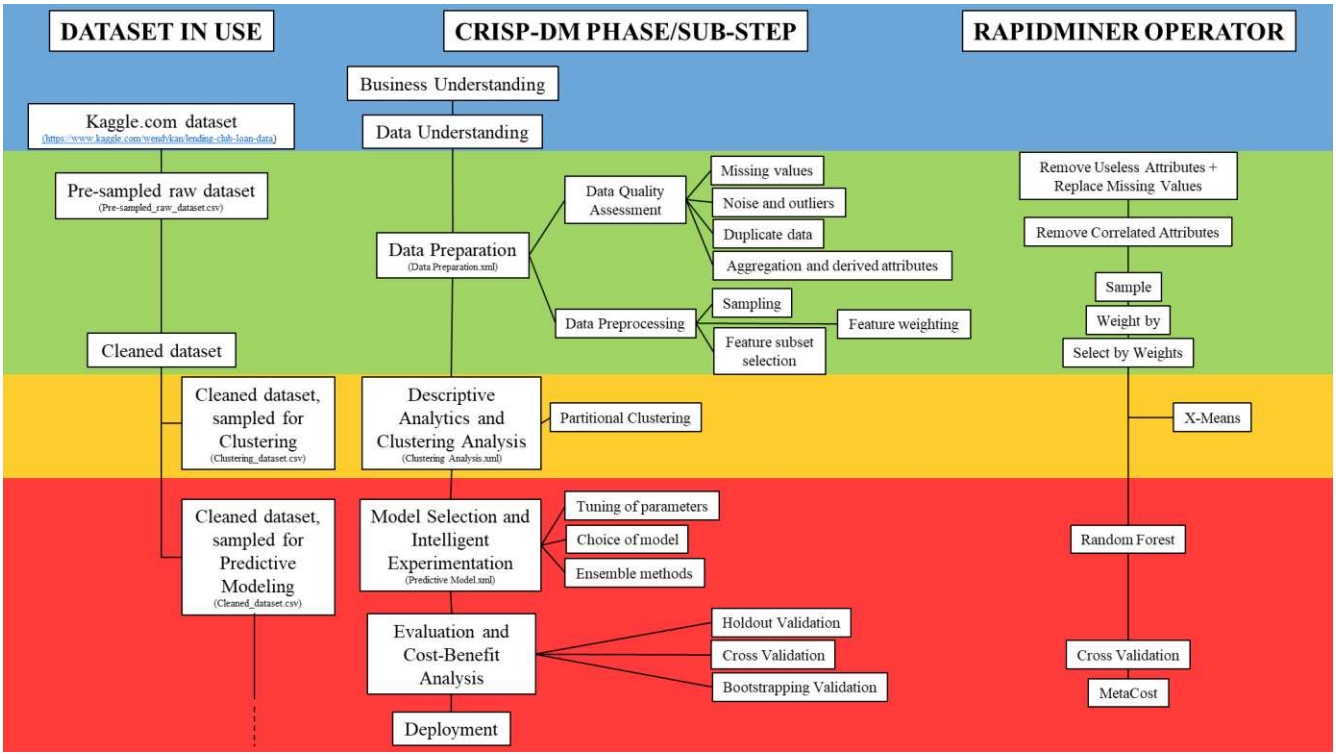
1. Each operator performs a specific and particular action on data.
2. Each operator contains precise input ports and output ports: connecting one operator to the next using one or more of them delivers some particular outcomes. The main ports that will be used are:
    a. Exa: It receives or returns a full table of the ExampleSet, which is a data table containing all records for each attribute at a given step in the analysis.
    b. Wei: It receives or returns an ExampleSet containing all attributes and an associated weight value, where each weight represents the feature importance for the given attribute.
    c. Clu: It receives or returns the outcome of a Clustering model.
    d. Mod: It delivers the outcome of a specific model, trained on the whole ExampleSet.
    e. Tra: It receives the training set.
    f. Tes: It delivers the test result set in the form of an ExampleSet.
    g. Per: It delivers the performance of the model over a given ExampleSet.
    h. Res: It allows for the visualization of the final outputs of the whole analysis process.

3. Different operators contain distinctive parameters either to be tuned by hand by the analyst or to be optimized with the use of another dedicated operator.

Finally, the data table at hand takes on the name of ExampleSet in RapidMiner, and it is composed of attributes and examples. Attributes (also referred to as features or variables) are represented by columns in the data matrix, while examples are the rows.

The following pages will guide the reader through the creation of a Clustering Analysis model and a Predictive Model, both trained on the same LendingClub loan dataset that will be presented later on.

CRISP-DM: CRoss Industry Standard Process for Data Mining [(10)] As the name itself says, this framework is a standard procedure to be followed whenever it comes to facing a data mining problem. CRISP-DM was developed by a group of data scientists at SPSS and Daimler-Chrysler in the late 1990s, and it is the most popular framework for analytics, as its high-level focus makes it generalizable to most data mining tasks. The framework is composed of the following six phases (plus one, specific to our case, meaning "Clustering Analysis"), each of them containing specific sub-steps:



To be noticed how in the diagram right above, we have put right next to each other all the CRISP-DM phases, the relevant operator to address them in RapidMiner, and the development of the dataset at hand across the project. This is a very general outline of the empirical part of the work, an overall view of what will happen in the following pages: different-colored bands visually separate the different macro-

phases that make up for the whole bigger flow necessary before reaching the final deployment of the models and its results. The goal from now on will be to extensively go through each phase, describing which challenges from the problem at hand it specifically tries to assess, and how. Right after outlining the issues to be addressed, the practical demonstration of what can be done to solve them in RapidMiner will be outlined.

Business Understanding As the very first phase of the whole process, this step takes care of defining the goals for the analytics project from a business perspective, a task that is attained also by delineating the setting and the project requirements, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

Important questions that could be useful in guiding the analyst through this phase include: Who is your client? Which problems are more important for him? Which problems are more suitable to be addressed with analytics? What are the analytics goals and how could they solve the identified problems? What are the success criteria? What are the KPIs?

Assessing the situation is crucial at this stage of the process, as it allows to highlight resources, constraints and assumptions in general terms, but very early on, channeling the work on the desired track. Additionally, a description of the intended plan for achieving the data mining goals (and thereby achieving the business goals) is due: the plan should specify the steps to be performed during the rest of the project, including the initial selection of tools and techniques.

Everything that relates to this phase of the framework was already presented extensively either in the theory section or in the introduction to the empirical section, so we will move on to the next step.

Data Understanding The data understanding phase starts with an initial data collection (and their integration, if multiple sources are involved) and proceeds with activities that enable the analyst to become familiar with the dataset, identify raw data quality problems, discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information. After listing the dataset(s) acquired, together with their locations and the methods used to acquire them, it is important to examine the “gross/surface” properties of the collected data and report on the results.

Relevant questions for this step include: Do we have the right, sufficient data? Which are the relevant attributes? Which is our target variable?

Exploring data comes right next: this task addresses data mining questions using querying, visualization, and reporting techniques. These include distribution of key attributes (for example, the target attribute of a prediction task) relationships between pairs or small numbers of attributes, results of simple aggregations, properties of significant sub-populations, and simple statistical analyses. These analyses may directly address the data mining goals, contribute to or refine the data description and quality reports, and set the ground for other data preparation steps needed for further analysis. Simple visualization (not necessary inclusion in the report) of graphs and plots to grasp data characteristics is very important at this stage of the whole process.

In the context of the RapidMiner project at hand, the dataset that is being used is public and taken from Kaggle.com, so not covered by NDA, and available at https://www.kaggle.com/wendykan/lending-club-loan-data . Data refers to the complete activity of LendingClub over the period 2007-2015, and it is thus a very rich and exhaustive source of information for some insightful analytics. However, as any kind of analytics project, also this work has to deal with an important tradeoff: representativeness of findings vs. time consumption vs. technical constraints. Obviously, the optimal situation would be to stand in the position that allows for the maximization of the first and the simultaneous minimization of the other two, but in reality this is very rarely attainable, also in the professional setting of big corporations. This is very important in the sense that the dataset available at the link accounts for as much as 145 variables (called "attributes" in RapidMiner) and over 2.26 million rows, each corresponding to a specific loan issued by LendingClub across the specified period. As a consequence of this, imagining to perform interesting calculations on such an ExampleSet while using a regular home computer is very difficult, especially considering some very practical facts relating to technical constraints (as we are not using a "super-computer") and time consumption (as we shall see later on, building the right model for the data requires going through a process consisting of several iterations and recording of performance, which is just unthinkable if every model, especially the more complex ones, requires as long as one hour to run due to dataset size). A consequence of this is that the original dataset downloadable at the link above went pre-sampled, so that its size got reduced to a more manageable 887,233 rows and 74 attributes. In the process, special attention was devoted to the preservation of representativeness of the sample, mainly thanks to random selection of instances, while at the same time bringing benefits from the point of view of time required for each calculation and less strain on computer's RAM, opening up to the use of more sophisticated models.

Even if sampling was beneficial from the point of view of time consumption and fitting to technical constraints, the newly-obtained dataset (Pre-sampled_raw_dataset.csv) is far from being usable in practice if one aims at obtaining precise results: this is mainly due to the fact that the ExampleSet at hand is still, in a certain sense, "raw". More precisely, at this point there are a number of issues with the data, some of them concerning the quality of the ExampleSet, in particular: a high number of missing values (to the point that some attributes are almost completely empty, thus useless), noise, outliers, and instances or variables being duplicate (or almost duplicate) of each other. Getting a sense of what issues hide amongst the dataset is crucial at this time, as it allows to select the right RapidMiner operators to solve them during the Data Preparation step of the CRISP-DM framework.

In addition, the Data Understanding step is very important in the context of the whole work also because it is the moment during which it is decided which deliverables will have to be prepared in light of the project submission. For this work, the additional elements (provided upon request) that will come together (as a unique zipped file) with this text as supplemental material are:

1. The pre-sampled, raw dataset before it entered the Data Preparation step, in .csv format (Pre-sampled_raw_dataset.csv);
2. The RapidMiner analysis process used to conduct the Data Preparation step and change the data at hand from the pre-sampled raw dataset to the cleaned dataset, in .xml format (Data Preparation.xml)
3. The cleaned dataset after the Data Preparation step and used for the Clustering Analysis, in .csv format (Clustering_dataset.csv);
4. The cleaned dataset after the Data Preparation step, sampled once more to be used for the Predictive Modeling, in .csv format (Cleaned_dataset.csv);
5. The final, optimized version of the RapidMiner analysis process used to conduct the Clustering Analysis, in .xml format (Clustering Analysis.xml);
6. The final, optimized version of the RapidMiner analysis process used to conduct the Predictive Modeling, in .xml format (Predictive Model.xml);
7. The Excel file containing both the Centroid Table for the Clustering Analysis, and the work of Intelligent Experimentation for the optimization of the Predictive Model, in .xlsx format (Clustering Analysis output & Intelligent experimentation for Predictive Model.xlsx);

8. The Excel file containing the data dictionary of the attributes used in the Clustering Analysis and Predictive Model, along with a brief explanation of what they refer to, in .xlsx format (Data dictionary.xlsx).

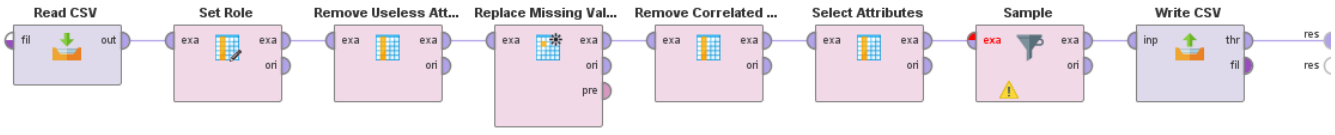
Data Preparation How should we fix the data quality issues found in the previous phase? How should we avoid sample causally? These are in fact two very relevant questions: out of the big set of data that we have, we want to make sure we apply the right techniques to end up with a "data lake", instead of a "data swamp". The Data Preparation phase covers all activities needed to construct the final, cleaned dataset (meaning those data that will be fed into the modeling tools) from the initial raw data: the central idea here is to raise data quality to the level required by the selected analysis techniques. As a consequence, this is an essential phase for achieving satisficing results in the next phases: it articulates in two sub-steps (Data Quality Assessment and Data Preprocessing), and it ends with a summary description of the cleaned dataset.

Data Quality Assessment takes care of handling and solving the data problems found in the previous phase. Main techniques involved include addressing:

1. Missing values: They are generated thanks to information not collected or attributes not applicable to all cases of the ExampleSet. This issue is solved either by elimination of specific data instances/attributes, or by estimation. Useful RapidMiner operators, then, are "Remove Useless Attributes" and "Replace Missing Values".
2. Noise and outliers: While the first issue refers to an unpredictable modification of the actual data and it is thus hard to address, the latter concerns the presence of instances with characteristics that are considerably different than most of the other instances in the ExampleSet. This second issue can be solved by simply removing the outlier values with the operators "Detect Outlier (Densities)" and then "Filter Examples".
3. Duplicate data: The dataset at hand may include instances or attributes that are duplicates (or almost duplicates) of one another, a characteristic that is assessed in terms of how high their correlation is. The issue is solved by removing attributes that are correlated above a given threshold by using the "Remove Correlated Attributes" operator.

Data Preprocessing, instead, is aimed at reducing the data volume (without losing informativeness) to match the technical computational constraints of the machine on which analyses are being performed. Main techniques involved are:

1. Aggregation and derived attributes: This step refers to combining two or more attributes (or objects) into a single attribute (or object) in order to reduce dataset size, change scale to some attributes, and reduce variability. Also, some other possibly relevant attributes may be derived from the ones already in the dataset.
2. Sampling: While statisticians sample because obtaining the entire set of data of interest is too expensive/time consuming, data miners sample because processing the entire set of data of interest is too expensive/time consuming: sampling then is often necessary in data mining due to computational/processing constraints. When sampling, it is crucial to balance representativeness with critical mass constraints: machine learning models are incentivized to learn patterns that apply to large groups, in order to become more accurate. Therefore, if a particular group is not well represented in the data because sampling eliminated most of its instances, the model will not prioritize learning about it, and results may come out biased. From here, then, we recognize the importance of sampling randomly (not causally) and considering the performance of a balanced vs. unbalanced sampling, which are all matters that can be assessed with the "Sample" operator.
3. Feature weighting: The idea of this step is to understand which features are more important for the determination of the outcome variable by picking them on the basis of how much they reduce the entropy of the dependent variable: a ranking is drafted, and features are attached a weight based on how much less "confusion" they add. If the dependent variable is nominal, this step is addressed with the "Weight by" operator, which comes in two main criteria Information Gain and Chi Squared.
4. Feature subset selection: Whenever the dimensionality of the dataset increases (mainly because many attributes are being considered), the volume of the multidimensional space increases so fast that the available data become too sparse, which is problematic for statistical significance: as a consequence, performing statistically sound analyses implies gathering and integrating more data, which is often very expensive. This phenomenon is called "Curse of Dimensionality", and it highlights the need to select just the right number of attributes, which in theory should be the most important ones, a task that is addressed with the "Select by Weights" operator.



Right above it is reported the screenshot of the RapidMiner process that has been built for the Data Preparation step of this work: the process takes as input the pre-sampled, raw dataset (Pre-

sampled_raw_dataset.csv) and outputs the cleaned dataset in .csv format thanks to the "Write CSV" operator. Specifically to the process above, the sample at hand gets treated by multiple operators, which in sequence: remove attributes who contain more than 80% missing values (and are, then, useless), replace the remaining missing values by estimation, remove attributes that are more than 90% correlated with each other, sample the data to create a more manageable ExampleSet (sampling type: absolute, balanced by "loan_status", total size equal to 14,136 examples). The problem of outliers has not been addressed in the above process due to the high time consumption of using such an operator in light of the technical means available. An additional remark on feature weighting and feature subset selection: these steps were included directly in the Predictive Modeling process, as they become handier in that context.

Descriptive Analytics and Clustering Analysis This step in the work refers to inspecting the existing data and describing them (not predicting something from them) in order to gain useful and interesting insights and explaining underlying actionable patterns. Therefore, differently to what it is required for a Predictive Analytics model, there may be no target/dependent variable. Moreover, this step overlaps with the more classical Business Intelligence (BI), as traditional BI tools are used. On this regard, there are a number of interesting descriptive analytics techniques, some of them being very trivial but necessary to get a grasp of the data at hand:

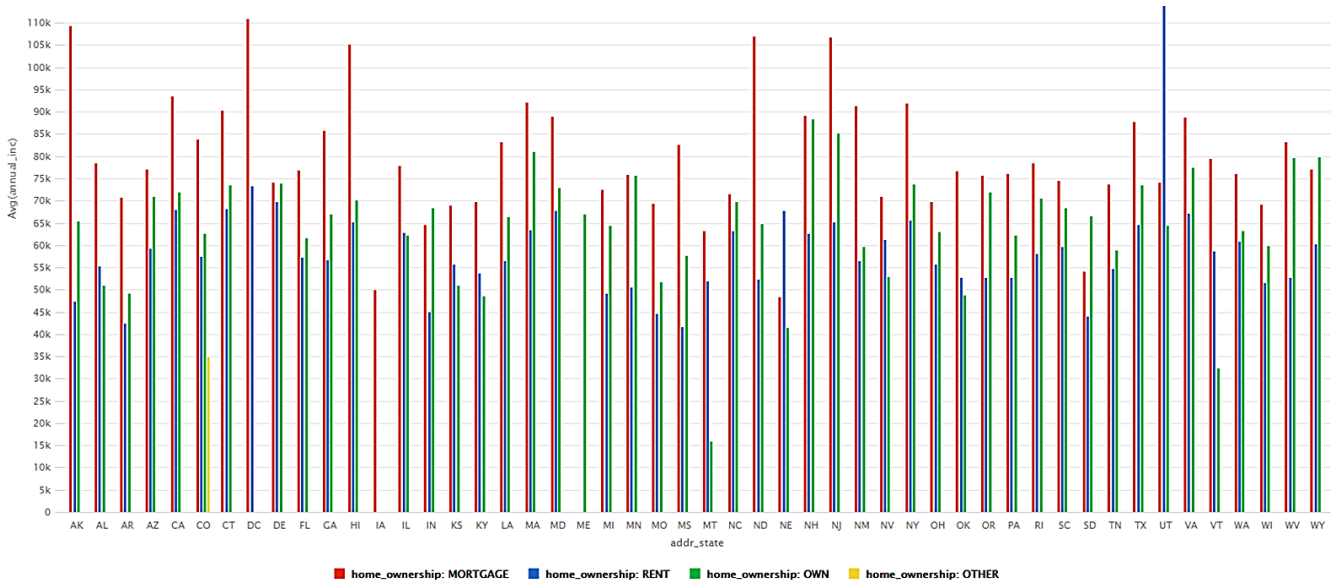
1. Descriptive statistics
2. Information visualization (charts, plots, …)
3. Descriptive modeling (Association Rule Mining; Clustering Analysis; Social Network Analysis)

Considering the specific setting of this work, the following descriptive statistics and information visualizations refer to the cleaned dataset after the Data Preparation step outlined in the previous pages.

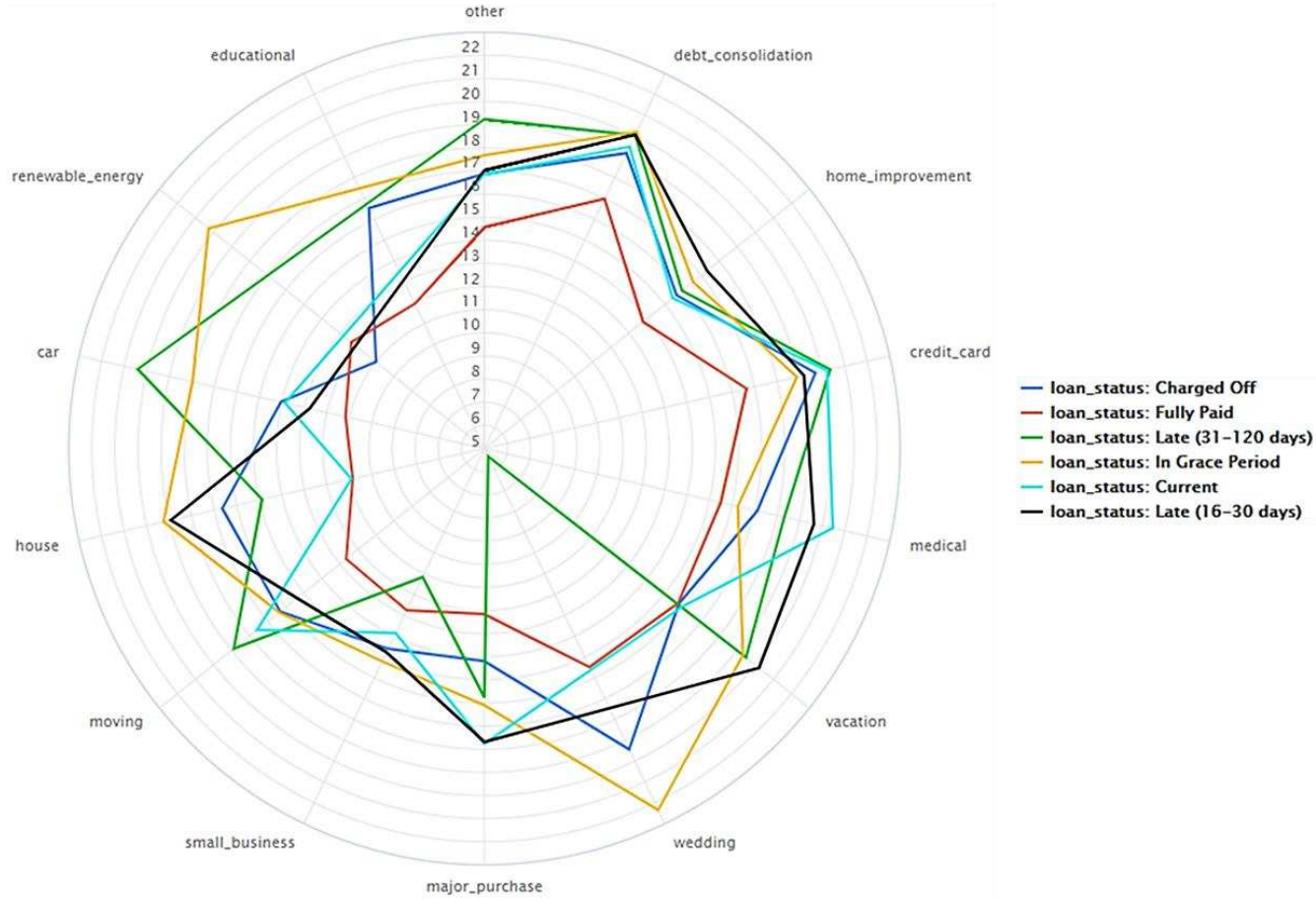
SUMMARY STATISTICS OF THE MOST RELEVANT ATTRIBUTES

ATTRIBUTE	TYPE	MIN/LEAST FREQUENT (IF NOMINAL)	MAX/MOST FREQUENT (IF NOMINAL)	AVERAGE/AVERAGE FREQUENCY (IF NOMINAL)	DEVIATION/FREQUENCY DEVIATION (IF NOMINAL)	VALUES (IF NOMINAL)
addr_state	Nominal	Maine (1)	California (2,214)	283	388	50 US States (including District of Columbia (DC), exluding Idaho and US Territories)
annual_inc	Real	$ 1,170	$ 2,146,496	$ 72,187.46	$ 49,274.71	
emp_length	Nominal	9 years (576)	10+ years (4,397)	1,217	1,078	< 1 year; 1 year; 2 years; 3 years; 4 years; 5 years; 6 years; 7 years; 8 years; 9 years; 10+ years
home_ownership	Nominal	Other (1)	Mortgage (6,592)	3,534	3,322	Mortgage; Rent; Own; Other
earliest_cr_line	Nominal	September 1973 (1)	October 2001 (122)	27	27	Range: November 1955-September 2012
term	Nominal	60 months (4,615)	36 months (9,521)	7,068	3,469	36 months; 60 months
loan_amnt	Real	$ 1000	$ 35000	$ 15012.2	$ 8542.36	
int_rate	Real	5.320%	28.990%	14.878%	4.541%	
purpose	Nominal	Educational (7)	Debt consolidation (8,699)	1,010	2,318	Car; Credit card; Debt consolidation; Educational; Home improvement; House; Major purchase; Medical; Moving; Other; Renewable energy; Small business; Vacation; Wedding
loan_status	Nominal	Late (31-120 days) (2,356)	Current (2,356)	2,365	0	Charged off; Current; Fully paid; In grace period; Late (16-30 days); Late (31-120 days)
dti	Real	0%	1092.52%	18.66%	12.33%	
sub_grade	Nominal	G5 (22)	C3 (813)	404	271	Ranges: A1-A5; B1-B5; C1-C5; D1-D5; E1-E5; F1-F5; G1-G5
issue_d	Nominal	November 2007 (1)	October 2014 (745)	144	179	Range: July 2007-December 2015
out_prncp	Real	$ 0	$ 35,000	$ 7,653.69	$ 8,343.97	
total_pymnt	Real	$ 0	$ 54,025.42	$ 7,816.5	$ 7,610.06	
recoveries	Real	$ 0	$ 18,173.53	$ 151.02	$ 692.52	

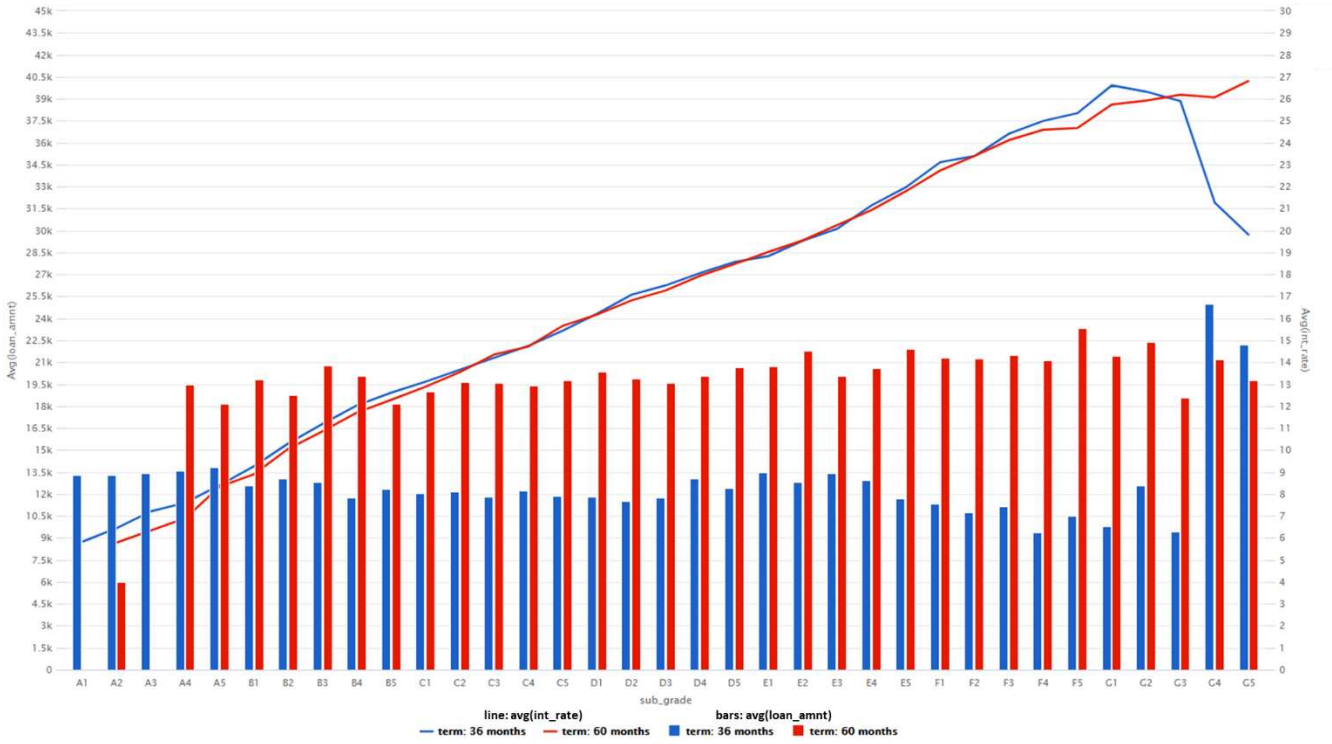
AVERAGE ANNUAL INCOME BY US STATE, BY HOME OWNERSHIP STATUS



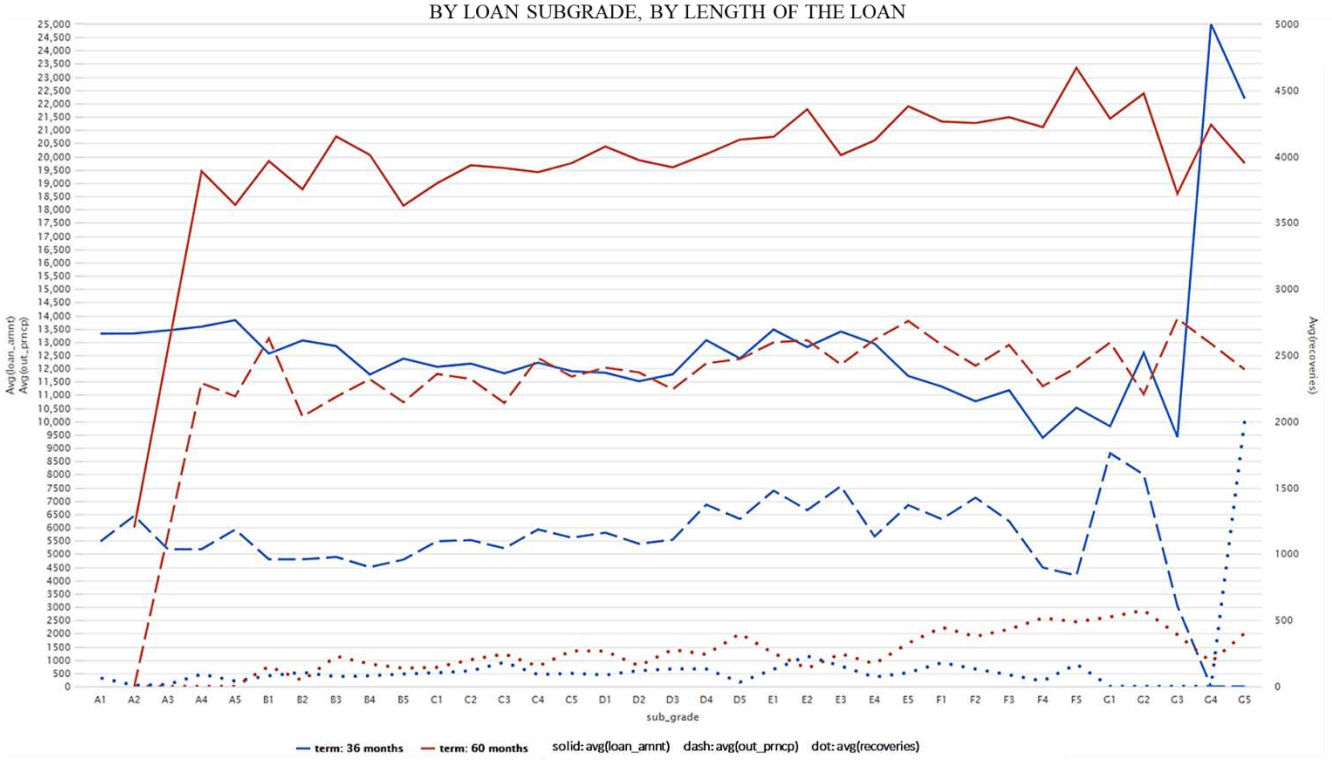
AVERAGE DTI RATIO BY LOAN PURPOSE, BY LOAN STATUS



AVERAGE LOAN AMOUNT AND INTEREST RATE, BY LOAN SUBGRADE, BY LENGTH OF THE LOAN



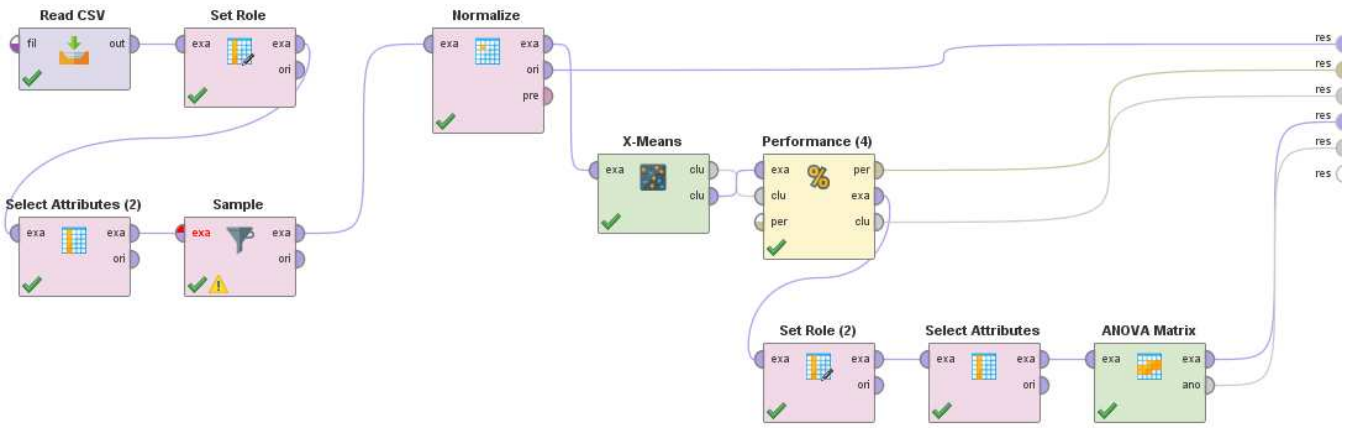
AVERAGE LOAN AMOUNT, OUTSTANDING PRINCIPAL, AND RECOVERIES (IF LOAN IS CHARGED-OFF): BY LOAN SUBGRADE, BY LENGTH OF THE LOAN



Instead, when it comes to descriptive modeling, again, there are many ways to gain insights from the data. The most useful one for the purposes of this work is "Clustering Analysis". This technique is based on finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups: mathematically, the idea is to minimize intra-cluster distances, while at the same time maximizing inter-cluster distances. An important specification is, however, necessary: Clustering Analysis requires no prior knowledge of cluster membership when classifying new cases, and grouping is not a result of an external specification (thus we do not give an explanation for why the cluster exists). It is crucial to understand the difference here, with respect to both Predictive Modeling (where we have class label information and we try to predict new unseen cases) and to simple segmentation of instances based on some trivial rule (for example, clusters obtained from alphabetical ordering, as they do not shed light on hidden patterns in the data other than pure randomness).

Furthermore, it is important to consider how there is not a unique clustering method: for our purposes, "Partitional Clustering" proves best, as data is divided into non-overlapping subsets (our clusters) such that each data object is in exactly one subset (therefore there is no hierarchy of clusters or nesting). The algorithm for Partitional Clustering is "K-Means", where k refers to the number of clusters we are aiming at. The choice of k can be performed either by selecting a pre-determined number of clusters (if prior knowledge about the dataset was available), or by trying out a range of k's (if no prior knowledge about the dataset). In this second case, the best k is the one which (as said before) minimizes intra-cluster distance while maximizing inter-cluster distance.

Specifically to our setting, the RapidMiner process implemented during Clustering Analysis is the one that follows:



The Data Preparation step to obtain Clustering_dataset.csv from the cleaned dataset is not present in the above screenshot, as it has been performed separately. While most operators used were the same ones that have been implemented also for the creation of the Predictive Model dataset (and with all their parameters tuned the same way), one more step has been necessary here: since the "X-Means" operator (with measure type set to numerical) requires only numerical attributes as inputs to run smoothly, all the relevant nominal attributes in the ExampleSet went through a process of binarization, performed with the "Nominal to Numerical" operator. Moreover, the Clustering_dataset.csv file has been randomly sampled to 100,000 examples, in order to meet technical and computational constraints. Finally, normalization of all attributes in the ExampleSet has been necessary in this case to avoid the situation in which cluster creation by "X-Means" gets biased by the different scale of variables (e.g. variables accounting for years span larger ranges than dummy variables, and not normalizing would then introduce implicit weighting of some features of the ExampleSet).

Going back to the above-pictured process, it delivers three main outputs, which allow the analyst to get a grasp of the quality of the clusters and get insights on what makes them different from each other. First of all, the main criterion to evaluate the explicative power of clusters is called "Performance Vector", a table that records the values of average intra-cluster distance for each cluster generated with the analysis process. These numbers are negative in the sense that they quantify "the cost of categorizing some elements within the same cluster", and the smaller they are in absolute value, the higher the Clustering quality is. The Performance Vector for our analysis is the following:

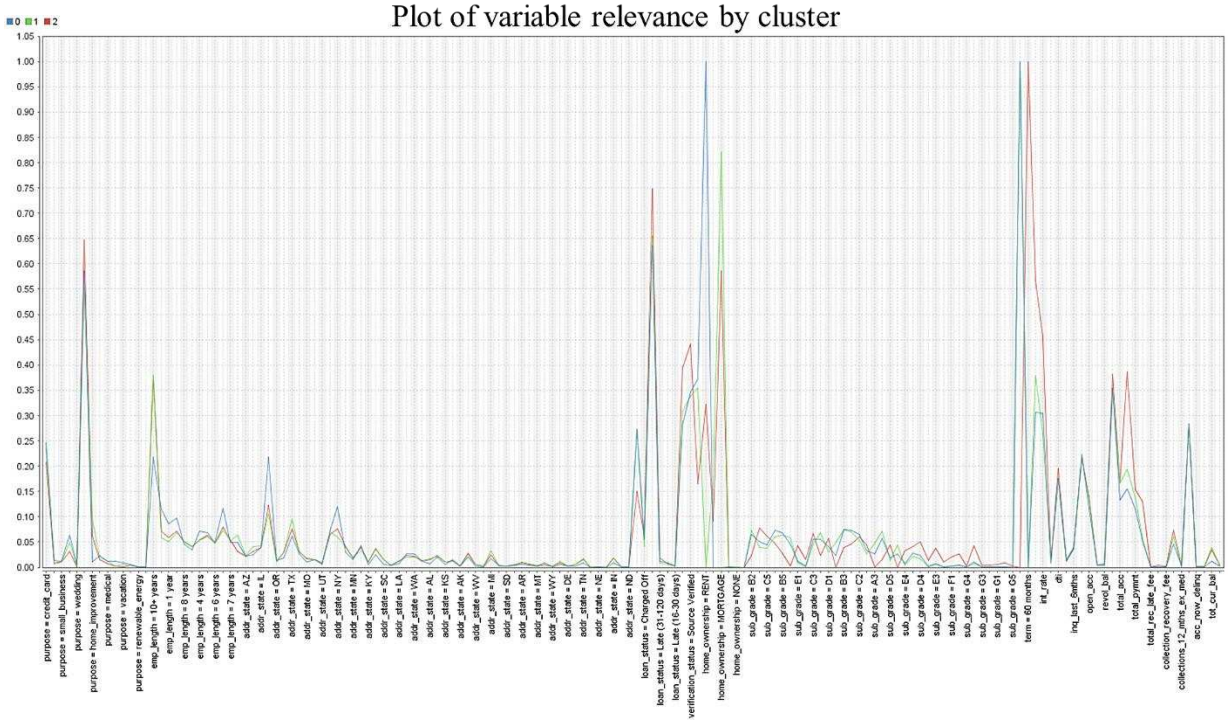
PERFORMANCE VECTOR	
Average within-centroid distance: Overall	-4.950
Average within-centroid distance: Cluster-0	-4.696
Average within-centroid distance: Cluster-1	-5.019
Average within-centroid distance: Cluster-2	-5.112

As we can see, RapidMiner suggests that our dataset at hand can be described with the use of 3 clusters, but we want to go further and inspect other features:

1. Which attributes of the ExampleSet are the most important ones in describing differences across clusters? Meaning, which variables have been the most influential ones in the creation of those different clusters?
2. How can the average member of each cluster be described? Meaning, which values does each attribute of the ExampleSet attain on average in each different cluster?

The first question is addressed with the use of the "ANOVA Matrix" operator, which performs and "ANalysis Of the VAriance" across the different clusters: this is a general technique that can be used to test the hypothesis that the means among two or more groups are equal, under the assumption that the sampled populations are normally distributed. The output of this operator is a list of p-values, one for each attribute: attributes that report a p-value which is larger than the selected type-I error (in our case, 0.05 and 0.01 significance levels were used) are those for which it holds that differences in the observable average value across clusters are not significant at the selected significance level. It follows, then, that those attributes are not relevant for the creation of meaningful different clusters (which is the ultimate goal of Clustering Analysis), and should, thus, not be considered when explaining both why clusters differ, and what characterizes the average member of each cluster.

The second question above, instead, is addressed by looking at the output of the "X-Means" operator, called "Centroid Table", which lists the average value that attributes attain within each cluster. In this sense, this table allows to get a grasp of the average persona representing each cluster, therefore highlighting average differences among the groups. Graphically, the "Centroid Table" is depicted using the following plot: all attributes are recorded on the x-axis, and each cluster is represented using a different-colored line, therefore higher peaks show more relevant attributes by cluster.



To be noticed is how the insights coming from the outputs of "X-Means" and "ANOVA Matrix" have been combined into a single Excel table (contained in the "Clustering Analysis output & Intelligent

experimentation for Predictive Model.xlsx" file) to make the comparison of clusters and determination of personas easier. On this regard, the 3 personas that can be identified in the dataset at hand are:

1. **Cluster-0:** Accounting for 30.12% of the whole sample, this group is characterized for having its members mainly residing either on the East Coast (New York (11.9%), New Jersey (4.3%), Pennsylvania (2.9%), Massachusetts (2.6%), Virginia (2.7%)) or on its Western counterpart (California (21.8%), Washington (2.6%)), with a significant share being from Texas (6.1%). Members of this cluster are polarized in terms of working history, with a major share having been employed for 10 or more years (21.8%), and another big percentage on shorter lengths (3 years (9.7%), 2 years (11.7%), 1 year (8.6%), for less than 1 year (11.4%)). Additionally, the almost totality of these customers already pays a rent for the house they currently live in, and has an average annual self-reported income of $ 61,531. After joining the LendingClub network, people of this cluster usually ask for a 36-months loan, of an average amount of $ 11,072 at average interest rate of 12.5464%, with main purposes for borrowing being debt consolidation (58.7%), credit card (24.7%), or need to afford a major purchase (2.3%). Historical data highlight how members of this group mainly have their loan status current (63.5%), already fully repaid on time (27.3%), or charged off (5.6%). In addition to that, on average, they have a DTI ratio of 17.41%, 34.7 months have passed since the last time they have been financially delinquent, and have a credit history that accounts for 21.49 credit lines granted (10.42 of which being currently open). Considering the rating of the loans that have been given, this group is mainly comprised of average-grade loans (B3 (7.4%), C1 (7.4%), B4 (7.2%)), while at the same time accounting for 17.9% of A-rated ones and as much as 31.3% in the E-F-G-rated band. The average post charge-off gross recoveries for members of this cluster accounted to $ 30.62.
2. **Cluster-1:** Accounting for 39.69% of the whole sample, this group is characterized for having its members mainly residing in the South-West (California (10.5%), Texas (9.6%)), East Coast (Florida (6.9%), New York (6%), Georgia (4%), Pennsylvania (3.9%), New Jersey (3.3%), North Carolina (3.1%), Virginia (2.9%)), or Mid-West (Ohio (3.8%), Michigan (3.3%)). In terms of working history, 38% of people in this cluster have been employed for 10 or more years, while a more uniform distribution is observable on the other employment lengths (approximately 5% of the members each). Additionally, borrowers in this cluster have an average annual self-reported income of $ 81,544, and 82.1% of them is paying a mortgage for the house they currently live in, while another 17.8% share owns their house. After joining the LendingClub network, people of

this cluster usually ask for a 36-months loan, of an average amount of $ 13,557 at average interest rate of 11.5917%, with main purposes for borrowing being debt consolidation (55.4%), credit card (23.4%), or home improvement (9.6%). Historical data highlight how members of this group mainly have their loan status current (65.5%), already fully repaid on time (27.2%), or charged off (4%). In addition to that, on average, they have a DTI ratio of 17.57%, 33.6 months have passed since the last time they have been financially delinquent, and have a credit history that accounts for 26.7 credit lines granted (11.8 of which being currently open). Considering the rating of the loans that have been given, this group is mainly comprised of higher/average-grade loans (B3 (7.6%), B2 (7.3%), A5 (7.1%)), while at the same time accounting for 26.9% of A-rated ones and 14.3% in the E-F-G-rated band. The average post charge-off gross recoveries for members of this cluster accounted to $ 27.02.

3. **Cluster-2:** Accounting for 30.19% of the whole sample, this group is characterized for having its members mainly residing either on the East Coast (New York (7.7%), Florida (6.6%), Pennsylvania (4%), New Jersey (3.9%), Georgia (3.2%), Virginia (3.2%), North Carolina (3%), Maryland (2.8%)) or the South-West (California (12.4%), Texas (7.7%)). Considering employment history, while 37.7% of members have been working for 10 or more years, other employment lengths account for 7.5% (shortest lengths, meaning less than 4 years) and 4.5% (longer lengths, meaning more than 4 years) of the cluster. Additionally, borrowers in this cluster have an average annual self-reported income of $ 80,916, and a very diverse home ownership status: 58.7% of them is paying a mortgage for the house they currently live in, 32.3% of them pays a rent, while another 9% share owns their house. After joining the LendingClub network, people of this cluster usually ask for a 60-months loan, of an average amount of $ 20,057 at average interest rate of 16.1012%, with main purposes for borrowing being debt consolidation (64.8%), credit card (20.9%), or home improvement (6.1%). Historical data highlight how members of this group mainly have their loan status current (74.9%), already fully repaid on time (15%), or charged off (6%). In addition to that, on average, they have a DTI ratio of 19.49%, 33.9 months have passed since the last time they have been financially delinquent, and have a credit history that accounts for 27.1 credit lines granted (12.3 of which being currently open). Considering the rating of the loans that have been given, this group is mainly comprised of lower/average-grade loans (C4 (7.9%), C3 (6.6%), C5 (6.2%)), while at the same time accounting for 26.9% in the E-F-G-rated band and as low as 1.9% of A-rated ones. The average post charge-off gross recoveries for members of this cluster accounted to $ 79.36.

Model Selection and Intelligent Experimentation We are now entering the heart of the thesis work. If so far data were just being explored, the time is right to put them to some good use in order to predict about the future, which is the core of our Analytics efforts. Important questions coming up at this stage are then: How could we find the best model? Should we even rely on a single, best model?

In this phase, various modeling techniques are selected and applied, and their parameters calibrated to optimal values. At the end of this phase, after one (or more) best-performing models have been selected, the analyst needs to document the actual modeling technique that is to be implemented in the end, along with specific assumptions made to run such model.

As the goal for the Predictive Model is to most accurately forecast previously unseen instances of the outcome variable, a specification on the methodology is needed. First of all, the target variable needs to be declared with the use of the "Set Role" operator: from now on, the specific role of class label is attached to the variable, turning it into a "special attribute" of the ExampleSet. In order to perform efficient modeling and validate findings, the labeled data is divided into training set and test set, such that the former is used to build a Predictive Model, and the latter to determine its accuracy. More specifically, the goal for the modeling phase is (by working on the training dataset) to find a model for the class label as a function of the values of other attributes (or independent variables), such that the precision with which the class label in the test set is predicted is maximized. A comparison of the predicted class labels with the actual class labels in the test dataset is what allows to determine the accuracy of the Predictive Model.

Since we are considering the prediction of the credit risk associated with customers of LendingClub, the target variable for this phase is "loan_status", a nominal categorical variable: this variable accounts for when the customer is expected to repay the loan that was granted to him. In particular, this variable takes on 6 different values depending on the time for a repayment by the borrower:

1. Current: The loan is outstanding and is currently being repaid on schedule.
2. Fully paid: The borrower has fully repaid the loan according to the scheduled terms.
3. In grace period: The borrower is late on a payment in a window of 0-15 days.
4. Late (16-30 days): The borrower is late on a payment in a window of 16-30 days.
5. Late (31-120 days): The borrower is late on a payment in a window of 31-120 days.
6. Charged off: The borrower is so late on a payment (more than 120 days) that the charge-off procedure for his loan is initiated, as there no longer are reasonable expectations of receiving

future payments as agreed. In addition, LendingClub proceeds with the collection of recoveries on the loan.

The ultimate objective for this phase of the work is, by training a Predictive Model on historical data by LendingClub, to most accurately predict what kind of customer the new borrowers applying for a loan will be, in terms of repaying according to the scheduled terms. In practice, the goal of the Predictive Model is to assign each new customer asking for funds to one of the above loan status categories, such that a better informed decision can then be taken when it comes to actually approving his grant. It is then clear how the more accurate the Predictive Model is, the better it is at predicting the loan status of each new customer, which in turn translates into a remarkably valuable tool in support of the decision to accept such borrower in the LendingClub network, along with reduction of wasted funds on borrowers that will repay late.

The sample at hand for this phase is the one obtained after the Data Preparation step (Cleaned_dataset.csv), which has been sampled to correct for its high imbalance in terms of the target variable (sampling type: absolute, balanced by "loan_status", total size equal to 14,136 examples). Not only that, but before running the model, the dataset went through sampling once again, and this time the reason is more bound to common sense rather than technicalities. Since customers of LendingClub are granted a 15-day grace period to make payments with no penalty [(11)], it is highly possible that many borrowers take advantage of this possibility, which makes the distinction line between clients with "Current" loan status and "In Grace Period" loan status extremely blurred. As a consequence, to avoid an unwanted biasing of the Predictive Model, borrowers with loan_status equal to "In Grace Period" have been excluded from the final dataset entering the modeling tools, which is exactly the role performed by the second sampling mentioned above, performed by the "Sample" operator included in "Predictive Model.xml". The final sample size is then 11,780 examples across 33 attributes in total.

At this point, everything is set for a solid Model Selection phase: the analyst starts by picking a type of Predictive Model, running it on the sample at hand and recording its performance in an Excel spreadsheet. Afterwards, he manually modifies the model parameters and sees what the effects on accuracy is. Additionally, the "Optimize Parameters" operator could come handy at this stage, as it allows to automatically find the best values for the selected parameters, even if not all of them at the same time: this is due to computational constraints, as the computer needs to solve an optimization process in several dimensions, considering all the different combinations achievable with the selected set of parameters to

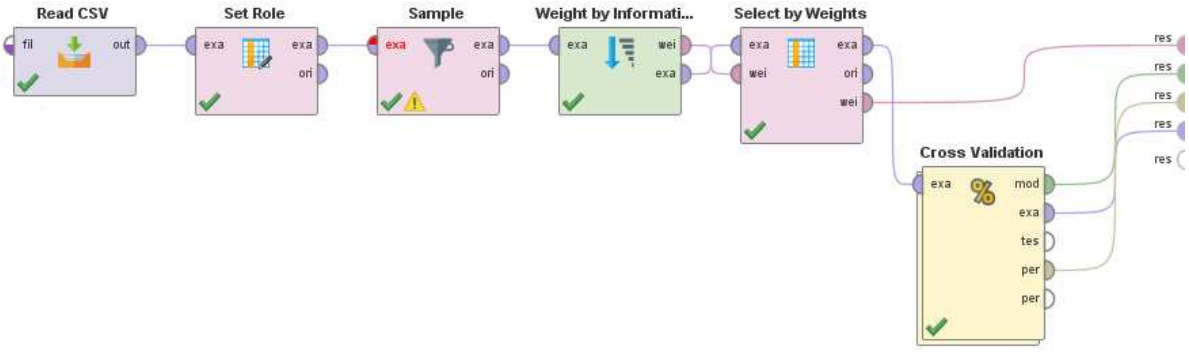
be perfected. Considering this, the initial manual adjustments to at least find a range over which the parameters perform better is then justified.

In tweaking parameters, we need to be aware of additional theoretical factors, such as achieving a balance between complexity and overfitting: in fact, an efficient Predictive Model reaches optimal complexity while maintaining reasonable generalizability, which translates into not overfitting to the training data. This is another important tradeoff to be taken into consideration, as when complexity increases, the model is more likely to overfit to the data; while if it decreases, the model is more likely to have high training errors (meaning that it is not sophisticated enough to capture the underlying patterns in the data). Considering how model performance varies when tweaking parameters turns out to be, again, very important, but it is not the end of the story: in fact, if two models of different complexity perform the same in terms of maximizing the accuracy on the test set while minimizing the spread in accuracy between the training and test set, the optimal model to be selected shall be the simplest one. This task could also be addressed in graphical terms with the Receiver Operating Characteristic (ROC) curve.

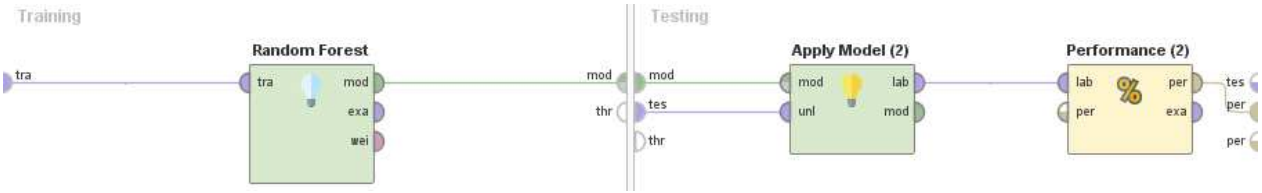
However, it does not stop here, as also different models (not only parameters) could be employed, and their performance recorded, to see if better results can be achieved. This whole process of trial, error, and gradual optimization is called "Intelligent Experimentation", in the sense that the analyst needs to experiment with many different combinations of models and parameters, applying his own knowledge of the theory and modeling skills to increase model accuracy the most.

Additionally, on top of trying out usual modeling techniques one at a time, they can also be combined with ensemble methods. The basic idea behind ensemble methods is that of constructing a set of models (which are called "classifiers" in this specific setting) from the training data by either: manipulating the training examples, manipulating the input features, or manipulating the learning algorithm. The prediction of unseen cases is then undertaken using the insights from different classifiers at the same time, combined by means of simple voting or by using more complex approaches. The advantage of using ensembles is clear and evident, and it mainly concerns risk mitigation: considering each classifier has a given error rate, if we assume each classifier to be independent from each other, the overall combined probability of predicting wrong is sensibly reduced.

Below it is reported a screenshot of the RapidMiner process of the best-performing model for the problem at hand, which is composed of two nested processes. The outer process is the following:



The inner one, instead, is nested under the "Cross Validation" operator, and it is as follows:



What goes on in practice when implementing the above Predictive Model in RapidMiner is that the analysis process takes as input the cleaned dataset after the Data Preparation step (Cleaned_dataset.csv) and performs some operations, which include setting "loan_status" as class label ("Set Role" operator), and sampling the dataset to remove "loan_status == "In Grace Period"" ("Sample" operator). After that, the "Feature weighting" step of Data Preparation is performed by using the "Weight by Information Gain" operator: the goal is to determine which ones are the most relevant attributes of the ExampleSet by calculating a weight for each of them using the Information Gain criterion. Right after, the "Feature subset selection" step of Data Preparation is carried out by means of the "Select by Weights" operator: specifically to the above analysis process, it is found to be optimal to keep the top 75% of attributes of the ExampleSet in terms of associated weight (of course, the higher the weight of an attribute, the more relevant it is considered to be in determining the value of the target variable). Next, the dataset finally enters the modeling tools: nested under the "Cross Validation" operator (whose role is assessing the performance of the model, and which will then be introduced in the next section), we find the Predictive Model that has led to the best performance out of the many trials and variants attempted, "Random Forest". This Predictive Model is actually an ensemble method, in the sense that what it does is building multiple Decision Trees (a type of Predictive Model), each one on a randomly selected subset of features, and delivering a final prediction that is made by combining the insights gained from individual trees by means of voting.

But what is a Decision Tree, the basic classifier behind the Random Forest ensemble method? Decision Tree is a type of modeling tool, possibly the most basic one, yet proving to be extremely handy and

applicable in a big number of cases. The name refers to the fact that, by its very structure, it closely resembles a tree, in the sense that they are both composed of branches and leaves. The idea behind this algorithm in the context of a Classification problem (thus when the outcome variable is categorical) is constructing a tree-like collection of nodes intended to make a decision on values affiliation to a class. In practice, what the algorithm does is recursively dividing the training set by means of sequential steps, represented by nodes in the tree: each node stands for a splitting rule for one specific attribute, and the creation of new nodes is repeated until some stopping criteria are met. One possible stopping criterion can be node pureness: for each branch of the Decision Tree, whenever at a specific node the algorithm reaches a division that is either pure o relatively small, it stops splitting further, and those final divisions at the final node take on the name of "leaves" of the tree. A pure division is, then, one for which each leaf contains only members of the ExampleSet from the same class. But then one natural question that should come up at this point is: How does the algorithm decide which attributes to split on to reach pureness the fastest? Actually, the algorithm reasons is in terms of entropy at each node to assess its pureness: the lower the entropy of a node, the higher the probability that it is selected to split the data. Obviously, minimization of the entropy (also referred to as Information Gain) is not the only splitting criterion that can be used in Decision Trees, but it is the most relevant to our context as it allows to be consistent with what has been chosen across the rest of this work.

Already from the description above, it is evident how this kind of Predictive Model is extremely versatile, and even more so if many different Decision Trees are implemented at the same time, which is exactly what the "Random Forest" ensemble method does. Not only that, as another advantage of using multiple trees to improve classification is that this helps to avoid building models that are overly dependent on only a few attributes.

A specification on the parameters used to reach the best performance in the context of the problem at hand with Random Forest include:

- Number of Decision Trees built: 200
- Splitting criterion: Information Gain
- Maximal depth for each branch of a Tree: 50
- Apply pruning (confidence = 0.001): This parameter reduces the size of the Decision Trees by replacing with leaves (according to the confidence parameter) branches of the tree that provide

little power to classify instances, thus lowering the complexity of the final classifier, while at the same time improving the predictive accuracy by the limiting overfitting.

- Apply pre-pruning (minimal gain = 0.0001; minimal leaf size = 2; minimal size for split = 2; number of pre-pruning alternatives = 3): This parameter specifies if more stopping criteria other than the maximal depth should be used during generation of the Decision Trees.
- Voting strategy to combine insights from different Decision Trees: Confidence vote.

This is the final, best-performing model that was found after the "Intelligent Experimentation" of a big number of models and combinations of parameters. The next section of the work will take care of walking the reader through how model performance is measured, along with how accurate the above-described model, implemented on the LendingClub data, turned out to be.

Evaluation and Cost-Benefit Analysis At this stage in the project, we have built a model (or models) that appears to have high quality from a data analysis perspective. Before proceeding to its final deployment, it is important to thoroughly evaluate model performances: but how should we evaluate the analytics outcome? How do we know that the results we obtain from our model are generalizable to other datasets and unknown future instances?. If previous evaluation steps only took care of considering factors such as the accuracy and generalizability of the model, this one looks at them altogether while also assessing the degree to which the model meets the business objectives and seeking to determine if there is some business reason why it is deficient. After evaluating models with respect to business success criteria, the most compliant one(s) becomes the approved model(s).

If up to now we have given for granted that the models automatically spit out a value of accuracy, this is certainly not what happens in practice, where we need to specify to the machine both how to evaluate its results and the proportions for how the ExampleSet has be divided between train and test set. These two factors have an enormous influence over the general performance indeed, as we are selecting how the algorithm has to validate its findings. Of course, then, there exist multiple validation methods to approach these tasks:

1. Holdout Validation ("Split Validation" operator): This method simply splits the ExampleSet into training and test data according to a proportion set by the analyst. It is the least computational expensive method, but also the most biased one.
2. Cross Validation ("Cross Validation" operator) [(12)]: With this method, data is partitioned into k disjoint random subsets, so that the model is trained on the first k-1 subsets and tested on the

remaining one. It is characterized by having the highest variance among all methods (thus it is upward biased), and it is then best with large datasets as variance gets naturally reduced.

3. Bootstrapping Validation ("Bootstrapping Validation" operator) [(13)]: This method is basically a Holdout Validation repeated k times using subsampling replacement. While it excels for being the method with the lowest variance, it is the most computational expensive one, thus not very suitable for large datasets.

Moreover, one notable fact is that, within the selected methods for data validation, the selected sampling type is the "stratified sampling". This method builds random subsets of the data and ensures that the class distribution in the subsets is the same as in the whole ExampleSet. For instance, in the special case of a binominal classification of the target variable, stratified sampling would build random subsets such that each subset contains roughly the same proportions of the two values of class label.

For the scopes of our work, we put together the "Cross Validation" operator and the stratified sampling feature, obtaining an evaluation procedure for which k random partitions (called folds) are created, such that the model is trained on the first k-1 folds and the last one is the test set. In the context of this work, 10 folds was selected as the optimal number of partitions. Another advantage that supports selecting the Cross Validation method is the fact that a confidence interval for model accuracy is returned, providing additional insights on the quality of the model performance value.

The output of the Evaluation step is called "Confusion Matrix", a data table that summarizes the predictions results in percentage terms for each value of the target variable.

CONFUSION MATRIX

		ACTUAL CLASS LABEL		
	Client repays?	**True class label = Yes**	**True class label = No**	**Class label precision (p)**
PREDICTED CLASS LABEL	**Predicted class label = Yes**	TP	FP (Missing target)	$Precision_{(TP)}$
	Predicted class label = No	FN (False alarm)	TN	$Precision_{(TN)}$
	Class label recall (r)	$Recall_{(TP)}$	$Recall_{(TN)}$	

Right above it is reported a Confusion Matrix for a prediction model where the class label is binomial (thus we get a 2x2 matrix), a simplified setting that is useful to show the important features of this object. First of all, contrarily to what has been considered so far, we immediately notice how more than one metric for performance evaluation could be looked at: this is the case if we want to go a step further,

recognizing how model accuracy might not always be the best metric for assessing the performance of a Predictive Model. In fact, there may be instances in which data could deliver better accuracy by not applying any model at all (Accuracy Paradox), and this would be completely useless in practice. Therefore it follows how, in order to avoid such a scenario, we need additional metrics to evaluate a Predictive Model. Some useful ones are:

1. How much the model predicts right overall: Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
2. Exactness of prediction for each predicted class label: Precision (p) = $\frac{TP}{TP+FP}$ *or* $\frac{TN}{TN+FN}$
3. Completeness of the prediction for each actual class label: Recall (r) = $\frac{TP}{TP+FN}$ *or* $\frac{TN}{TN+FP}$
4. Balanced measure of exactness and completeness: F-measure (F) = $\frac{2 \cdot r \cdot p}{r+p}$
   a. For TP: $F_{(TP)} = \frac{2 \cdot TP}{2 \cdot TP+FN+FP}$
   b. For TN: $F_{(TN)} = \frac{2 \cdot TN}{2 \cdot TN+FN+FP}$

Specifically to this work, the evaluation metrics that will be considered when assessing different models are overall accuracy and a product of all the recalls for the class label at hand (loan_status). The Confusion Matrix generated with the best-performing Predictive Model (optimized Random Forest model) described in the previous sections is as follows:

CONFUSION MATRIX FOR BEST PERFORMING PREDICTIVE MODEL

Accuracy: 87.33% +/- 0.83%
Product of the recalls: 49.85%

		ACTUAL CLASS LABEL					
	When does the borrower repay?	True loan_status = Charged Off	True loan_status = Fully Paid	True loan_status = Late (31-120 days)	True loan_status = Current	True loan_status = Late (16-30 days)	loan_status precision (p)
PREDICTED CLASS LABEL	Predicted loan_status = Charged Off	2287	91	1	1	1	96.05%
	Predicted loan_status = Fully Paid	69	2265	5	2	5	96.55%
	Predicted loan_status = Late (31-120 days)	0	0	1899	38	131	91.83%
	Predicted loan_status = Current	0	0	190	1959	341	78.67%
	Predicted loan_status = Late (16-30 days)	0	0	261	356	1878	75.27%
	loan_status recall (r)	97.07%	96.14%	80.60%	83.15%	79.71%	

As it can be seen by looking at the above table, the best-performing Predictive Model correctly forecasts the loan status of previously unseen borrowers (test set) 87.33% of the times on average (meaning overall, across all values of the class label), and with a confidence level of +/- 0.83%. Instead, the other metric considered (the product of all the recalls) is a useful one in the sense that it allows to simultaneously capture how accurate the model predictions are for each individual actual value of the class label, thus providing further insights on the quality of the model performance: for the above table, the product of the recalls is 49.85%. Especially notable is the performance for borrowers whose loan status is equal to "Charged Off" and "Fully Paid" (which in theory are the most important categories): the former

(borrowers who will be late on payments for more than 120 days) is correctly predicted 97.07% of the times, while the latter (borrowers who will fully repay the loan according to the scheduled terms) is forecasted right 96.14% of the times.

While the resulting overall accuracy of the Predictive Model may seem quite far from the typical industry standards, it is actually pretty remarkable if one thinks about the fact that this model performance has been achieved using just an ordinary home computer (this representing the main technical constraint of the project), relatively simple models (due to computational constraints of the machine at hand), and without the support of an Analytics team across all the phases of the work.

To the confusion matrix above, which defines how the model performs in technical terms over the data, another important matrix that has to be added to assess the business implications (advantages or not) from implementing the selected model, is the "Cost Matrix", which shows all the costs for the model predicting right/wrong with respect to actual values of the class label. Notice how negative costs in this matrix mean benefits of predicting right/wrong with respect to actual values of the class label: this is why this matrix is often referred to as "Cost-Benefit Matrix".

COST-BENEFIT MATRIX

		ACTUAL CLASS LABEL	
	Client repays?	**True class label = Yes**	**True class label = No**
PREDICTED CLASS LABEL	**Predicted class label = Yes**	Cost of TP	Cost of FP (Missing target)
	Predicted class label = No	Cost of FN (False alarm)	Cost of TN

Thinking specifically about the context of LendingClub and going back to a simplified 2x2 scenario, obviously the FP case is the most costly one (our model predicts the client will repay on schedule, thus we grant the loan, but he actually repays late), thus the one that has to be more precisely predicted in the confusion matrix. In purely statistical terms, in the context of the 2x2 confusion matrix, it is the type-I error of our model, which has to be minimized, while at the same time not forgetting about the presence of the type-II error (FN). However, things become more complicated when considering our actual setting, where we have to deal with a 5x5 confusion matrix, which therefore implies also a 5x5 cost-benefit matrix. It follows that focusing on asking the right question in the first place is the most important thing to avoid conducting a completely unsound Cost-Benefit Analysis. Specifically to the context of this work, the goal during this step of the CRISP-DM framework is not trying to give an answer to "What would have happened (meaning, how much would have been saved) if LendingClub had not granted the loan to

borrowers predicted to have a loan status of, for example, "Charged Off"?", but rather quantifying in budgetary terms how much the implementation of our predictive model would help LendingClub save. But saving with respect to what? This (apparently simple) question is in fact not trivial at all, as further pondering on how to practically give it an answer immediately opens up to one fundamental issue: up to now we have not considered what the current state of the art is at LendingClub when it comes to borrowers' credit risk scoring, which could be a good reference element to draw a meaningful comparison. In fact, assuming that the Cost-Benefit Matrix does not change after the implementation of the predictive model (meaning all costs associated with each prediction stay constant), seeing how beneficial/detrimental the application of our optimized model will be in terms of corporate budget saved in the future only entails a comparison with the performance achieved by the model currently in place, projected to the future. In symbols, the amount of corporate funds saved from implementing our predictive model with respect to the one now in place (S) is:

$$S = CB \cdot (C_{current\ model} - C_{our\ model})$$

where CB is the Cost-Benefit Matrix, $C_{current\ model}$ is the Confusion Matrix for the model currently in place at LendingClub when predicting the loan status of future new borrowers, $C_{our\ model}$ is the Confusion Matrix of our best-performing predictive model (reported in previous pages).

Therefore, what we need at this point is, first of all, LendingClub's current confusion matrix reflecting current predictive abilities with respect to previously unseen datasets. These information are usually easily available (in the company context) as part of the typical reports on future projections of corporate activities made by analysts, both at LendingClub as at any other company in the world. For our case, however, as we do not have such projections available that easily, it is mandatory that we come up with an alternate solution, a possibility being relying on comparing the performance of our predictive model to that of a benchmark one. A model that is very straightforward (and also easily implementable in terms of simulation) to be used as a benchmark for comparison with the one we previously designed is a "Random-Assignment" Predictive Model. In practice, this can be thought of as being a "non-model", in the sense that if LendingClub was to apply it for predicting borrowers' credit risk, it would be obtaining completely casual results, since what it does is attaching to each customer asking for financing a completely random loan status out of the ones available.

The Confusion Matrix for our benchmark model (Random-Assignment) is the following:

CONFUSION MATRIX FOR RANDOM-ASSIGNMENT PREDICTIVE MODEL

Accuracy: 19.70%
Product of the recalls: 0.03%

		ACTUAL CLASS LABEL					
	When does the borrower repay?	True loan_status = Charged Off	True loan_status = Fully Paid	True loan_status = Late (31-120 days)	True loan_status = Current	True loan_status = Late (16-30 days)	loan_status precision (p)
PREDICTED CLASS LABEL	Predicted loan_status = Charged Off	491	437	506	455	467	20.84%
	Predicted loan_status = Fully Paid	473	443	486	485	469	18.80%
	Predicted loan_status = Late (31-120 days)	491	505	441	445	474	18.72%
	Predicted loan_status = Current	451	470	481	477	477	20.25%
	Predicted loan_status = Late (16-30 days)	450	501	442	494	469	19.91%
	loan_status recall (r)	20.84%	18.80%	18.72%	20.25%	19.91%	

As expected, the performance of the Random-Assignment model is a very poor one indeed, with an overall accuracy of 19.70% and a product of the recalls equal to 0.03%. Just out of intellectual curiosity, it is still interesting to notice how, while these values highlight the fact that random assignment is certainly not a wise choice when it comes to predicting borrowers' expected loan status prior to granting them a loan, odds of getting it right by chance are still (incredibly) in the order of 1 out of every 5 clients.

Now that we have a baseline model that delivers predictions (easily obtained because actually no particular model at all was applied), it is time to make a very strong assumption regarding the context of this work, yet at the same time proving an extremely efficient move for simplifying the analyses: from now on, it will be assumed that the status quo at LendingClub is forecasting credit risk of new borrowers by using the Random Assignment Predictive Model. As said, this is just an assumption, and obviously it is very far from what the reality is at LendingClub, but for us it proves extremely handy to very evidently show how the implementation of an accurate and tailored Predictive Model (like the one designed in previous pages) can help a company save huge amounts of funds, which may then be re-directed towards other scopes and projects.

The problem has now been framed in such a way that meaningful and valuable findings can be obtained: by leveraging insights gained during the Descriptive Analytics phase, it is possible to estimate a 5x5 Cost-Benefit Matrix for the problem at hand, where the figures reported are the costs (benefits if negative values) deriving from predicting right/wrong each instance of loan_status against its actual value. In other words, the matrix that is obtained summarizes the net costs (or benefits) for each possible outcome of the prediction. Afterwards, by taking advantage of matrix multiplication, this constant Cost-Benefit Matrix gets multiplied by the difference of the two previously-found Confusion Matrices. What is then obtained is a breakdown of saved funds by loan status in the case of implementation of the best-performing Predictive Model with respect to the Random-Assignment baseline scenario, which then can be summed up to find the total amount of corporate budget saved.

The plan is clear, but there is still a missing piece: how is the Cost-Benefit Matrix calculated? As said, an estimation of the costs could be drawn by looking at the Descriptive Analytics phase of this work, leveraging the descriptive statistics and information visualization steps in particular. The goal from now on will then be to guide the reader through the process of estimation of the Cost-Benefit Matrix, step by step. Before doing so, however, some initial assumptions are needed in order to delineate the framework:

1. All values and costs that come are expressed in terms of averages (avg( )), as they will need to be multiplied by the confusion matrices at the very end
2. Recalling how across this work we are allowing only for credit risk and not default risk, a natural implication is that whenever we see a borrower not repaying (even if very late), we should assume that a payment is always made in the end. Also, at the time of repayment, the borrower gives back the whole amount funded by LendingClub (principal and interests), so that the case in which we might expect a borrower repaying only partly (default risk) is ruled out.
3. The calculation of interests (both on loans to borrowers and on re-investments of funds by LendingClub) is performed by means of simple interest, rather than by compound interest.
4. Headers for the matrices that will be shown have been excluded for brevity, but their ordering has been kept consistent (both on the x-dimension and the y-dimension) with the one in the previous tables, that is: “Charged Off”; “Fully Paid”; “Late (31-120 days)”; “Current”; “Late (16-30 days)”.

Diving right into the process for the calculation of costs, we start off by going back to the findings from Descriptive Analytics, in order to obtain, for each value of loan_status:

- The weighted average of the loan amount granted (avg(A)), weighted by loan subgrade and term, in dollars ($):

$$avg(A) = \begin{bmatrix} 14{,}585.19 & 14{,}585.19 & 14{,}585.19 & 14{,}585.19 & 14{,}585.19 \\ 13{,}553.28 & 13{,}553.28 & 13{,}553.28 & 13{,}553.28 & 13{,}553.28 \\ 15{,}461.84 & 15{,}461.84 & 15{,}461.84 & 15{,}461.84 & 15{,}461.84 \\ 15{,}047.92 & 15{,}047.92 & 15{,}047.92 & 15{,}047.92 & 15{,}047.92 \\ 15{,}673.95 & 15{,}673.95 & 15{,}673.95 & 15{,}673.95 & 15{,}673.95 \end{bmatrix}$$

- The weighted average of the interest rate offered on the loans (avg(i)), weighted by loan subgrade and term, in percentage (%):

$$avg(i) = \begin{bmatrix} 16.0689 & 16.0689 & 16.0689 & 16.0689 & 16.0689 \\ 13.3128 & 13.3128 & 13.3128 & 13.3128 & 13.3128 \\ 16.0768 & 16.0768 & 16.0768 & 16.0768 & 16.0768 \\ 12.8560 & 12.8560 & 12.8560 & 12.8560 & 12.8560 \\ 15{,}7007 & 15{,}7007 & 15{,}7007 & 15{,}7007 & 15{,}7007 \end{bmatrix}$$

- A time matrix (T_{Late}) reporting the difference (in months) between the actual time of repayment and the predicted time of repayment:

$$T_{Late} = \begin{bmatrix} 0 & -4 & -3 & -4 & -3.5 \\ 4 & 0 & 1 & 0 & 0.5 \\ 3 & -1 & 0 & 1 & -0.5 \\ 4 & 0 & 1 & 0 & 0.5 \\ 3.5 & -0.5 & 0.5 & -0.5 & 0 \end{bmatrix}$$

At this point, for each value of loan_status, we want to calculate the average predicted repayment (avg(R)) that LendingClub expects to get back when lending to a borrower:

$$avg(R) = avg(A) + avg(A) \cdot avg(i)$$

Also, we want to calculate the average actual repayment ($avg(R)_{actual}$) that LendingClub expects to get back when lending to a borrower:

$$avg(R)_{actual} = [avg(A) + avg(A) \cdot avg(i)] - [avg(A) + avg(A) \cdot avg(i) \cdot T_{Late} \cdot r] =$$
$$= avg(R) - avg(R) \cdot T_{Late} \cdot r = avg(R) \cdot [J - T_{Late} \cdot r]$$

where J is a 5x5 all-ones matrix. Also, r is the average interest rate that LendingClub hypothetically obtains when reinvesting repayments made by borrowers within its network in the form of other loans: obviously, this is a simplification of a much more complex element to be estimated, in the sense that it captures the opportunity cost of the money coming back on time to LendingClub in the form of repayment by borrowers. Mathematically, the calculation of average reinvestment rate (r) (expressed in percentage (%)) is an arithmetic average of the weighted averages of the interest rate offered on the loans (avg(i)), across all types of loan statuses (which are 5). In symbols:

$$r = \frac{\sum_{loan\_status} avg(i)}{5}$$

Moving on with the calculation of the costs, by taking the difference between the predicted and the actual repayment for each loan status, it is possible to isolate the component that refers to the amount of funds given up (in terms of missed opportunity for reinvestment) when LendingClub predicts a borrower repays

according to a given loan status, but in fact he actually ends up returning funds later/earlier than that. In other words, this new matrix (Ω) quantifies, in dollars ($), the cost of missed opportunities for reinvesting funds when the Predictive Model at hand forecasts wrong.

$$\Omega = avg(R) - avg(R)_{actual} = avg(R) \cdot T_{Late} \cdot r =$$
$$= \begin{bmatrix} 0.00 & -10{,}023.95 & -7{,}517.96 & -10{,}023.95 & -8{,}770.96 \\ 9{,}093.57 & 0.00 & 2{,}273.39 & 0.00 & 1{,}136.70 \\ 7{,}970.38 & -2{,}656.79 & 0.00 & 2{,}656.79 & -1{,}328.40 \\ 10{,}055.69 & 0.00 & 2{,}513.92 & 0.00 & 1{,}256.96 \\ 9{,}395.79 & -1{,}342.26 & 1{,}342.26 & -1{,}342.26 & 0.00 \end{bmatrix}$$

Looking at the above matrix, one fact should immediately catch the reader's attention, and not a minor one indeed: there may be instances across all the different combinations of predicted vs. actual loan status for which predicting wrong actually brings benefits (and not costs) in terms of opportunity cost of funds given up due to missed reinvestment opportunities. These cases (showing up as negative costs in the above matrix) actually pose LendingClub in front of a situation of moral hazard: giving up on reinvestment opportunities (of funds repaid by borrowers) proves beneficial if the Predictive Model over-estimates the delay in repayment with respect to the scenario that actually verifies. In other words, if we allow only for the existence of all those possible states of nature attainable within the boundaries set by the assumptions made at the beginning (regarding the calculation of costs and reinvestment possibilities), an extreme scenario would be LendingClub always predicting that borrowers will be repaying at the very last day available before becoming defaulting, so that it has the certainty that a repayment will actually come earlier than that, bringing benefits with respect to expectations. In this situation, LendingClub would be better off by purposely biasing the model to have it over-estimating the time before repayment, as opposed to predicting right, which would bring neither costs nor benefits in terms of reinvestment possibilities.

Not only that, but as of now the calculation of costs associated with wrong predictions is not taking into account the importance of different errors. For example, predicting that a borrower will be repaying on time and according to the scheduled terms while in fact he proves to be late by 120 days, is certainly a much more serious error if compared to that of predicting a borrower will be on time with repayments while in reality he turns out to be late by 16 days. To put it in terms of opportunity costs, in the first case the cost of missed opportunities for reinvesting funds is obviously bigger, and this because the time span over which interests from reinvesting could have been raised is longer.

The two problems that have just been highlighted with regard to the last matrix above can be solved with a unique solution: attaching a weight to each possible case. Of course, the weight does not have to be a random one, rather it should be 0 for the cases of moral hazard (so as to rule them out completely), and proportional to the seriousness of the prediction mistake (obviously in normalized terms) for all other cases. This weight matrix (W) is as follows:

$$W = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0.25 & 0 & 0.125 \\ 0.75 & 0 & 0 & 0.25 & 0 \\ 1 & 0 & 0.25 & 0 & 0.125 \\ 0.875 & 0 & 0.125 & 0 & 0 \end{bmatrix}$$

By multiplying the Ω matrix by the latter, we quantify, in dollars ($), the cost of missed opportunities for reinvesting funds when the Predictive Model at hand forecasts wrong, weighted to correct for moral hazard and seriousness of prediction mistakes: this is the actual Cost-Benefit Matrix (CB) for our work.

COST-BENEFIT MATRIX

		ACTUAL CLASS LABEL				
	When does the borrower repay?	True loan_status = Charged Off	True loan_status = Fully Paid	True loan_status = Late (31-120 days)	True loan_status = Current	True loan_status = Late (16-30 days)
PREDICTED CLASS LABEL	Predicted loan_status = Charged Off	0.00	0.00	0.00	0.00	0.00
	Predicted loan_status = Fully Paid	9093.57	0.00	568.35	0.00	142.09
	Predicted loan_status = Late (31-120 days)	5977.78	0.00	0.00	664.20	0.00
	Predicted loan_status = Current	10055.69	0.00	628.48	0.00	157.12
	Predicted loan_status = Late (16-30 days)	8221.32	0.00	167.78	0.00	0.00

At this point, we have all the elements we need to calculate how much implementing our best-performing Predictive Model allows LendingClub to save with respect to the Random-Assignment Predictive Model that has been assumed to be the status quo to assess borrowers' credit risk. A remark on this is that the following step can be automatically performed in RapidMiner (if the Cost-Benefit Matrix has been estimated and it is known) by using the "MetaCost" operator. Below it is reported the table for the total amount of funds saved (in dollars ($)), broken down by predicted loan status vs. actual loan status.

SAVINGS FROM IMPLEMENTING THE BEST PERFORMING PREDICTIVE MODEL IN PLACE OF THE RANDOM-ASSIGNMENT MODEL (S)

		ACTUAL CLASS LABEL				
	When does the borrower repay?	True loan_status = Charged Off	True loan_status = Fully Paid	True loan_status = Late (31-120 days)	True loan_status = Current	True loan_status = Late (16-30 days)
PREDICTED CLASS LABEL	Predicted loan_status = Charged Off	0.00	0.00	0.00	0.00	0.00
	Predicted loan_status = Fully Paid	3673801.19	0.00	273375.37	0.00	65928.36
	Predicted loan_status = Late (31-120 days)	2935090.69	0.00	0.00	270328.56	0.00
	Predicted loan_status = Current	4535117.81	0.00	182887.93	0.00	21368.35
	Predicted loan_status = Late (16-30 days)	3699593.16	0.00	30368.54	0.00	0.00

Taking the grand total of the last table ($Tot\ funds\ saved = \sum_{loan\_status} S$), we can exactly quantify the overall total savings deriving from implementing the best-performing Predictive Model designed over the course of this work with respect to the one assumed to be currently in place at LendingClub (Random-Assignment Predictive Model). The value found is exactly $ 15,687,859.95 saved, which (considering a sample size of 11,780 borrowers) amounts to $ 1,331.74 per person, or alternatively 8.96% of the average loan amount per person granted by LendingClub.

Deployment The creation of the model is generally not the end of the project, as the knowledge gained needs to be organized and presented in a way that customers, investors, and the management can use it: this is achieved by carefully crafting a final report and attaching all the required deliverables. Moreover, the Deployment phase takes care of identifying and proposing specific recommendations on how to improve and revise current corporate operations/processes in light of the insights derived from the analyses just performed.

CONCLUSION

Over the course of this work, we have hinted at why data and Business Analytics can be valuable tools in any corporate context nowadays, and additionally shown a very practical example of what it means to work with data for extracting profitable insights to enhance existing business practices and customers' satisfaction. All of this has been carried out in the context of peer-to-peer lending, taking into consideration the largest company in the industry and working with real-world data related to its activities over the period 2007-2015. The analyses performed were aimed at building from scratch two analytics models, one for Clustering Analysis of LendingClub's customers and one for Predictive Analytics of the credit risk assessment of future clients applying for a loan over the same platform.

In a nutshell, results achieved are remarkable. First of all, Clustering Analysis was able to give insights over the kind of borrowers applying for a loan through LendingClub by dividing them in 3 different average personas, who were described in striking detail: this could come very handy for tailoring initiatives and optimized lending conditions for new borrowing applicants, depending on the cluster they better fit into. Secondly, the proposed Predictive Model for the credit risk assessment of new borrowers eager to obtain financing thanks to LendingClub's network: the average accuracy of predictions across all possible categories of expected repayment was 87.33%, with a confidence level of +/- 0.83%. Not only that, as implementing this proposed model for credit risk prediction would have allowed LendingClub to save (with respect to a current benchmark model) exactly $ 15,687,859.95, amounting to $ 1,331.74 per person, or alternatively 8.96% of the average loan amount per person granted over the peer-to-peer platform.

These are impressive results, but is this all? Can we confidently say that modern technology has finally allowed humans to reach the "pot of gold", meaning a situation in which we can perfectly predict future outcomes, precisely optimize corporate processes, and definitively assess any situation of risk and uncertainty? For sure the answer is "no", and not just because, as we described in previous pages, our analytics capabilities are (as of today) inherently bound to the finite (although ever increasing [(14)]) computing power and speed. [(15)]In spite of this, however, rather than relying on human judgement alone, many organizations are increasingly asking for support from algorithms to weigh in on matters that potentially have profound social reverberations, for example whether to recruit someone for a job, give them a loan, or even identify them as a suspect in a crime. Although AI decision-making is often regarded as structurally objective, the data and processes that inform it can hide inequality behind apparent fairness, in the form of biases. Many articles present a number of machine learning biases, which either

come directly from the disposable data or that are sort of embedded in the applied models (see https://aibusiness.com/three-notable-examples-of-ai-bias/ for reference on this). Efforts to make analytics free of biases should be made both by producers of those models (through the adoption of transparent, accountable and carefully tested models against biases) and by the Governments: this has been the intent of the EU when it drafted and then enforced the GDPR regulation of 2018, detailing rules on the production and use of Analytics to make decisions while protecting citizens' right to privacy.

On the producers side of the fight against biases in Analytics, instead, particular attention should be placed on understanding the nature of the data, "consciously biasing" the models to counter-balance actual biases by weighting specific items/variables; or even removing the hard way protected class labels (for example "ethnicity") from the datasets. This last suggestion is a very common (and easier) way of mitigating biases in AI, as it only requires eliminating sensitive classes from the data, so that the model cannot learn from those and infer biased conclusions, which from its point of view would only be products of calculations and optimizations, thus rational. This approach would allow for a better control over AI biases in situations that involve, for example, criminal assessment and justice (see the COMPAS case for reference on this [(16)]). Additional ways of dealing and preventing biases in AI could also be of a more structural kind, for instance increasing diversity in AI teams, using de-biasing algorithms, and even setting up corporate-specific best practices for the application of Analytics solutions and their findings.

We are right in the middle of a turning point in history, as the AI revolution is something after which the world will hardly go back from, and is thus changing society and corporations forever. What is, however, important to always remark refers to how Analytics solutions are complementary, and not perfect substitutes, to human intervention and judgement. For instance, in the context of Credit Risk Analytics that has accompanied us throughout the whole development of this work, the AI software may be extremely good at predicting when a possible borrower will be repaying, thus informing the company on whether it is a wise choice to grant him the loan he is asking for. However, a machine (at least nowadays) is not yet able to discern on grounds different than calculations and optimizations (specific assumptions given): in simple terms, it lacks human judgement, which both in the context of granting a loan or in any other setting, is what gives humans the final say over anything that has to be done. The power of Analytics is a valuable tool that modern companies should take advantage from, as this is an unprecedented opportunity in history, but always keeping a fundamental idea in mind: a world where humans and machines work together, for sure is a place where either humans or machines working on their own will be outperformed.

WORKS CITED

1. O. LaBarre, *"What Is Credit Risk?"* (May 2nd, 2020), Investopedia (https://www.investopedia.com/terms/c/creditrisk.asp)
2. J. Kagan, *"Default Risk"* (May 19th, 2020), Investopedia (https://www.investopedia.com/terms/d/defaultrisk.asp)
3. TED, *"A smart new business loan for people with no credit | Shivani Siroya"* (May 18th, 2016), YouTube (https://www.youtube.com/watch?v=kSR8G8mfp84)
4. T. H. Davenport & J. G. Harris, *"Competing on Analytics: The New Science of Winning"* (2007), Harvard Business School Press
5. E. Stubbs, *"The Value of Business Analytics"* (December 2011), Analytics Magazine – INFORMS (http://analytics-magazine.org/the-value-of-business-analytics-2/)
6. T. Fountaine , B. McCarthy & T. Saleh, *"Building the AI-Powered Organization"* (2019), Harvard Business Review, July-August 2019 issue, 62–73
7. A. Shropshire, *"Using Machine Learning to Recommend Investments in P2P Lending"* (September 8th, 2019), Hacker Noon (https://hackernoon.com/using-machine-learning-to-recommend-investments-in-p2p-lending-cw2ll3yvh)
8. *"LendingClub"* (May 25th, 2020), Wikipedia (Wikimedia Foundation) (https://en.wikipedia.org/wiki/LendingClub)
9. C. Wilhelm, *"Google-Lending Club Alliance Takes Captive Finance into Digital Age"* (January 26th, 2015), American Banker (https://www.americanbanker.com/news/google-lending-club-alliance-takes-captive-finance-into-digital-age)
10. P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer & R. Wirth, *"CRISP-DM 1.0: Step-by-step data mining guide"* (1999), SPSS Statistics
11. "*Late Payments*" (2020), LendingClub (https://help.lendingclub.com/hc/en-us/articles/214575437-Late-payments#:~:text=That's%20why%20all%20our%20members,to%20cover%20the%20additional%20interest.)
12. B. Efron, *"Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation"* (1983), Journal of the American Statistical Association, 316–331
13. B. Efron & R. Tibshirani, *"Improvements on Cross-Validation: The. 632+ Bootstrap Method"* (1997), Journal of the American Statistical Association, 548–560

14. J. Desjardins, *"Visualizing Moore's Law in Action (1971-2019)"* (December 09th, 2019), Visual Capitalist (https://www.visualcapitalist.com/visualizing-moores-law-in-action-1971-2019/)

15. L. Hudson, *"Technology is Biased Too. How do we fix it?"* (July 20th, 2017), FiveThirtyEight – ABC News (https://fivethirtyeight.com/features/technology-is-biased-too-how-do-we-fix-it/)

16. E. T. Israni, *"When an Algorithm Helps Send You to Prison"* (October 26th, 2017), The New York Times (https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html)

17. M. North, *"Data Mining for the Masses"* (2012), Global Text Project

18. R. Mandelbaum, *"What Lending Club's Success Means for the Future of Small-Business Lending"* (April 13th, 2015), Inc.com (https://www.inc.com/magazine/201505/robb-mandelbaum/lending-club-money-on-demand.html)

19. P.-N. Tan, M. Steinbach & V. Kumar, *"Introduction to Data Mining"* (2006), Pearson Addison-Wesley

SOFTWARE

1. RapidMiner Studio (vv. 9.5) (https://rapidminer.com/)