Deep Science Publishing
https://doi.org/10.70593/978-93-49910-41-6_3

DeepScience
Open Access Books

Chapter 3: Machine learning in action: Predictive power in credit and risk models

3.1. Introduction to Machine Learning

Machine learning has, in recent years, become one of the most important developments in the world of technology and big data. Computers process and store large amounts of structured and unstructured data. Machine learning methods can allow computers to sift through and classify this data, recognize patterns, and make predictions or decisions based on those patterns. These methods can help in a wide variety of decision-making problems: they do not need detailed theory about the underlying process, and they can identify subtle and complex interactions. Cutting through all the hoopla and hyperbole surrounding machine learning, it is possible to identify a few simple use cases that account for most of the successful applications of ML. Generally, almost all machine learning revolves around four important steps: (1) choose representation or feature set, (2) choose a model, (3) choose a loss function, and (4) choose an optimization algorithm.

In financial services, the use of machine learning has become widespread. Consumer and commercial lenders, payment card issuers, investment banks, and asset management companies all use machine learning to some degree to answer a broad range of questions. These range from predicting which consumers are likely to pay back a loan, identifying which suspects are likely to be engaged in money laundering, recognizing which transactions are likely to be fraud, classifying consumer and small business loans that should be sold, trading on opportunities in the market, and optimizing consumer collections when a loan has gone into default. Companies that do not use advanced analytics methods like these in their risk and marketing operations have long been recognized as leaving money on the table. But today, many of the most sophisticated players are relying on ML not only to extract more data from their current operations but

also to do things that were impossible before because of the volumes of data that must be collected, processed, and understood to complete (Khandani et al., 2010; Lessmann et al., 2015; Bastani et al., 2021).

3.1.1. Definition and Scope

Credit is the cornerstone of today's financial system, providing access to capital to entrepreneurs, students, and families across the globe. Extending credit presents a significant risk to the lender or investor providing the funds. Typically, this risk is captured by default and loss models, which provide the probability or amount of default on a loan over a particular horizon. The development of a credit model requires many complex judgments, such as deciding on the type and amount of data to include and how this data should be utilized; determining the transformation, scaling, and interactions of this data; choosing the modeling framework; and developing the type(s) of stress tests. While there has been a lot of exploration in the selection of predictors to use in a credit model within the finance community, before the availability of more powerful automated machine learning packages, this work was largely informed by experience .

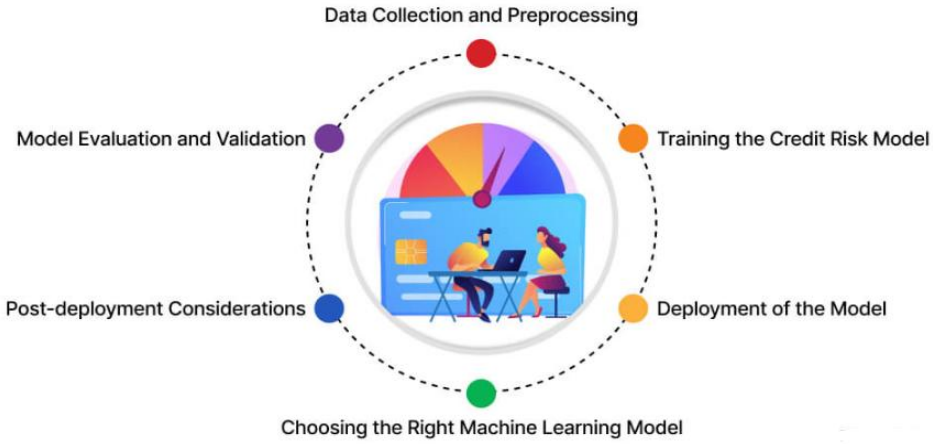

Fig 3 . 1 : Credit Risk Models with Machine Learning

To the best of our knowledge, there is no comparison of the relative predictive power of manual approaches and the newer automated and unsupervised ones on such a large and diverse sample of data. What is particularly interesting for us is the role that the unsupervised methods play as a filter for the final learning process. Unsupervised learning helps in reducing dimensionality and detecting patterns within a dataset, and can then pass this information to a supervised model. In the case of credit models, this can translate into great benefits if the analyst can focus their efforts on a smaller number of very relevant credit features, which in turn usually leads both to faster model

development iterations and to higher model performance (Malekipirbazari & Aksakalli, 2015; Sirignano & Cont, 2019).

3.1.2. Historical Context

Economists studied early warning models and developed theories such as the financial instability hypothesis and the debt deflation theory. The financial stability literature appeared much later, mainly motivated by the Asian Financial Crisis and then further reinforced after the Global Financial Crisis. Early warning models have been successfully applied for different purposes other than financial stability, such as predicting business cycles and detecting exchange rate crises. The existing literature identifies several leading indicators of systemic banking crises, such as financial liberalization, asset price bubbles, expansionary monetary policy, rapid growth of domestic credit, and external imbalances as a consequence of trade balances or total external debt, that may lead to financial crises.

The literature that explores the feasibility and effectiveness of early warning system techniques has gained momentum. Some of the existing studies apply these methods to banking crises in developed countries or more specific events due to urgent policy necessity. The identification of prudential policies to prevent systemic banking crises is still a challenging task due to policymakers' inconsistent policy responses resulting from a lack of information or the influence of powerful interest groups. These inconsistent policies exacerbate and aggravate the problems by softening the constraints and returning to market participants the distorted signal that there is little risk in such fragile and brittle systems.

3.2. Fundamentals of Credit Scoring

In response to the lack of control over the excessive granting of credit, large financial institutions created models based on artificial intelligence techniques, logistic regression, and artificial neural networks, among others. With these techniques, it has become possible to automate decision-making processes, analyze huge data volumes, and extract more value from them. In many sectors, these models became omnipresent, becoming a powerful weapon of competition, but also of social transformation. Credit risk problems have some characteristic aspects that make methods of statistical associative analysis well-suited to their analysis. One of the characteristics of the credit risk problem is the existence of a historical database, which links the default of debtors to some of their features, such as race, age, sex, salary, and education, among others. These features need to exhibit some degree of discrimination with respect to the precursors of default. From the statistical point of view, the most appropriate method is

one that predicts the probability of a certain event, given the presence of certain conditions. In finance, this method is known as the predictive method of regression.

3.2.1. Traditional Credit Scoring Models

Machine Learning in Action: Predictive Power in Credit and Risk Models

Statistical and machine learning models have a very specific goal when applied in the development of credit scoring models and the prediction of creditworthiness: to help identify current or prospective loan clients who are financially stressed and unable, or soon to be unable, to fulfill their contract obligations. To this end, in spite of the complex nature of the surrounding model and application processes, the primary output of a credit scoring model is a binary classification, i.e., a risk prediction that often considers and establishes lines in the sand for higher or lower affirmative and negative classifications. In credit scoring model work, the developer expects that above a given score threshold the borrower will repay as agreed and below that score he or she may experience a credit event or fail to meet his or her commitment.

How the borrower is classified is based on his or her pairing with a labeled data record, say a defaulted debtor or a performing loan. Will this borrower obtain a mortgage loan at a certain risk rate? Credit scoring models guide business decisions around lending or credit policies, providing powerful analysis tools across entire portfolios and in the granting of credit, all meeting the credit strategy goals of an organization. The straightforward structure and potential heavy transparency of traditional published scoring models have inadvertently misled many to focus too intently on the model structure, its rigidity, and its apparent simplicity. However, there is a black box effect in underestimating these models.

3.2.2. Limitations of Traditional Models

Wall Street demands models that can be automated and implemented within the financial institution's infrastructure and generate consistent and transparent evaluations and outputs across all the portfolios. While traditional models have delivered considerable predictive power to our clients, we have found them to be highly sensitive to different versions or releases of client data and input parameters. Data range adjustments, constant and default thresholds, industry defaults, and a mix of consulting time versus non-time series fundamental-based scores can result in vastly different financial statement models. Other model inputs such as assets under management, mutual fund ratings, number of employees, and tracking error can compound the volatility, since a financial company's answers based on our models are computed with these numbers.

The recent implementation of scoring model production guidelines has helped somewhat with these problems but has not totally explained or solved them. High levels of manual adjustment for integrating newer processing and underwriting techniques for criterion identification also underscore the limitations of traditional techniques for ultimate implementation. Holistic rules relating to a broker's identity or past successes in the underwriting of casino junk bonds are examples of some proposed effects we are simply told about anecdotally. For these and similar industry concerns, the structure and profile of the new generation of scoring models deserve exploration.

3.3. Machine Learning Techniques

Machine learning models often generate significant improvements in the predictive accuracy of credit and risk models. However, there are several important issues that complicate the use of a rigorous statistical learning approach to the estimation, including the issues of data relevance and data quality, as well as the trade-off between more statistical models and the very strong legal constraints imposed on credit and risk models. The chapter begins by introducing the notions of classification and regression, and then reviews some of the most important and successful statistical models used in credit and risk models.

Once a powerful model is selected on the basis of its ex-ante performance on the learner data, the model is laid for classification or estimation on some independent data to test its performance. Specifically, we explain some interesting machine learning algorithms, such as adaptive and computer adaptive learning, neural networks, boosting, support vector machines, and random forest, that are often used in the analysis of credit markets. Due to the complexity of credit and risk associated with prediction models, it is recommended that these models be used directly for decision making or for risk evaluation.

3.3.1. Supervised Learning

In the context of credit and risk models, supervised learning focuses on associations between features and labels that have been assessed by credit analysts or back-office employees. In a general form, supervised learning is defined as a process to identify a prediction mapping from features to labels. In a modeling context, the features are the matrix of features included in or derived for the model, while the labels represent the dependent variable whose potential relationships with features we are generally trying to identify. The most frequent focus is on estimating a model that provides a specific predictive accuracy benchmark in a 'live' creditor, marketing, or connected operations

setup. Model methodologies differ in aspects such as assessment data handling, sometimes creating temporary models purely for other models' input.

Conversely, similar modeling may appear in a variety of business processes, for economic, capital, or risk governance reasons, with model outputs driving processes not only associated with 'front line' credit or marketing deployment. Model outputs not only support capital computations, risk-adjusted product or customer pricing, and strategic marketing analyses, they may also carry back into credit loss or credit-related revenue models. While our table draws a strong focus on quantitative processes, we do not necessarily underestimate that model outputs require at least some approval process or ongoing back-calibration in a large number of instances, nor do we expect readers to agree with all our allocations between the depicted processes. At the same time, we deliberately did not seek to cover process aspects of model industry or academic image in the model portrayal.

3.3.2. Unsupervised Learning

Unsupervised learning is a very challenging area of statistics and machine learning. The main problem solved by unsupervised learning is to find categories or subgroups in complex data. There are generally fewer tools available for unsupervised learning problems since often a human can interpret the structures found in data or use domain-specific methods that don't rely on pure statistical methods. The main idea of unsupervised learning is to find some kind of structure in the data rather than relationships between variables. We can use unsupervised learning to understand the patterns in the data before applying more complex statistical methods. Often, the initial knowledge extracted via unsupervised learning is applied during data preprocessing.

The methods used in unsupervised learning are very different compared to those in supervised learning. Some of the main unsupervised learning techniques include clustering or finding subgroups of data, and finding subpopulations in mixture modeling or finite mixtures. Another common problem is mapping high-dimensional data into either fewer dimensions to gain insight or a larger, normally spatial dimension to examine the overall structure of the data in the context of unsupervised learning. Dimensionality methods such as principal components analysis and multidimensional scaling are often used for these tasks. Finally, square contingency tables are another form of unsupervised learning since they represent counts or frequencies of events and are used to examine the structural relationships in the data. This is an unsupervised method since no output is specified a priori.

3.3.3. Reinforcement Learning

Reinforcement Learning is learning what to do – how to map situations to actions – so as to maximize a numerical reward signal. Here's what it is relative to other forms of intelligent systems:

- Reinforcement Learning

- Reward and punishment

- Different competition

- Learning from interaction

- Temporal credit assignment

Reinforcement Learning is used for robots, video games, board game players, and other autonomous systems. These same methods often apply well to various applications and customer service. It carries its own terms and distinct assumptions which are totally unimportant in other intelligence work, but it also shares much with them. It has its unique take on the problem of credit assignment, which is about credit for a good action going to the earlier actions that deserve the most responsibility. In order to learn to act optimally and predict the consequences of actions that otherwise cannot be ranked, learning from reinforcement is one of the areas of reinforcement learning. Reinforcement can be designed to occur at the learning agent's discretion, but often comes from the agent's environment.

3.4. Data Preparation and Feature Engineering

Data preparation in the context of credit scoring is the process where a data scientist seeks to understand the data, detect and fix errors, and define how to use the data for the modeling process. It includes the creation of processes that are used for the initial development of a statistical model and in its deployment. An important sub-task within this process is feature engineering: the construction and transformation of covariates to provide additional information to help the model learn the outcome we want to forecast. The last part of this process is to define how the development process will produce data that is used in the modeling process – by defining treatments that will be applied to the production pipeline. In this chapter, we will examine these tasks and present potential answers to the challenge of preparing data for the implementation of credit scores.

In this section, we describe at the practical level a process for how to prepare data in the context of retail credit risk. In order to do this, we will use a data set that contains portfolio and loan-level information for a large set of loans. These loans are bad credit

risks, and the creation of a scoring model to differentiate them has inherent challenges. Because the loans are risky, the data presents a diverse set of profiles, some of which may not even resemble the demographic information relevant to the customer population used for acquisitions.

3.4.1. Data Collection

Introduction In the stress of consumer debt, credit models should play a central role—models that, to some extent, use the characteristics of each individual customer. These or 'behavioral' characteristics are defined by standard payment information. They are typically enriched in stress periods with additional information, the understanding of which can contribute to a more detailed risk and payment model. Data Collection After checking for moral hazard or the home-made lemons problem, we included 32,982 contracts representing 24,040 usable customers. This mitigates adverse selection or fake customers to maintain the positive selection bias of the data set. Monitoring more than 100 variables, including the scoring values, standard banking information, payment history, other existing contractual obligations, demographics, real estate information, job situation, and history, we distinguish and collect more knowledge than other researchers on the basis of the following data items of a given individual.

3.4.2. Data Cleaning

Once you've squared away your data sources and gathered your data, it's time to give your data some love and attention to ensure everything is 'clean.' This may include removing or correcting mistakes, standardizing formats, and making sure everything is keyed with a unique, distinguishable identifier. Some steps in data cleaning are common to many datasets, but sometimes they are specific to a particular cleaning task. Although cleaning data doesn't make for very exciting reading, without a relatively clean dataset it is almost impossible to make any headway in a modeling task.

Data for credit risk modeling must be scrubbed and engineered in a thoughtful way. The historical data will undoubtedly contain relevant variables that could possibly be leveraged to improve the modeling exercise. In the same token, the historical data will undoubtedly contain mistakes, missing values, and pooled incumbent information at different points in time. In the real world, data shows up with quirks and errors that need to be accounted for or mathematical solutions employed. This is how the real world works, and tackling the unpredictable is what makes the work interesting and rewarding. Complete historical data is a blessing, and more significant data from the distant past is a bonus!

3.4.3. Feature Selection

A credit risk or bankruptcy model will ideally use the best input data available to forecast the likelihood of an event in the future, that is, to deliver a high level of predictive power. However, feature selection is difficult to achieve when employing many data variables simultaneously. Unlike a model dealing with a simple target variable where the data scientist can simply look at different input variables one by one and compare some statistic, default or bankruptcy forecast models require simultaneous accounting for a large number of input variables, thus training a model using multiple features. However, not all input variables or features actually bring any useful information to the model, and, by definition, irrelevant information is worthless. This is why banks run into issues with regulatory tests about their compliance as the regulator will question the input variable selections and the resulting accuracy the institution obtained.

One approach is to use especially interpretable models, or black-box models such as decision trees that have feature evaluation built into them. Another option is to use specifically dedicated feature-selection techniques. These usually work by training a model on one set of input variables and then evaluating the resulting performance. The worst features are then foregone, and the exercise is repeated until only the key variables remain. Such feature-selection procedures can be helpful, but more information is indeed being used to reject variables that bring no new information. Of course, the features foregone may still contain some interesting information; thus a subsequent decision could actually be reducing the predictive power of the model.

3.5. Model Selection and Evaluation

Model selection is a fundamental part of predictive analysis using machine learning. In financial modeling, how to evaluate and select models is an ongoing question because models are often complex. Many financial models depend on data that are generated or affected in similar ways. The resulting correlation between the modeling data and the evaluation data undermines traditional statistical techniques.

Once data have been collected, prepared, and explored, the next step is to select the predictive model that will be used to generate predictions. There are a variety of predictive models, each with different strengths and weaknesses. The choice of model depends on the goals of the analysis and the nature of the inputs. For example, linear regression may be used for explanatory modeling with a continuous dependent variable. Classification and regression trees may be used for an analysis that requires easy interpretation of the model, and support vector machines may be used when good prediction performance is more important than model interpretation. Regardless of the model chosen for prediction, it is likely that there are variations in the input data that

will enhance the performance of a given algorithm and that some other variations will detract from its performance. Therefore, the creation of predictive models that perform well under a wide array of structural input changes requires an understanding of lead-lag relationships between financial and macroeconomic data.

3.5.1. Common Algorithms

There are a handful of different machine learning approaches that have proven to be extremely effective as tools in predictive credit models in the context of predicting default rates, delinquency rates, and loss given default. It is important to have a good understanding of the nuances of these techniques before using them in the context of credit risk. We divide these broadly into three different types: Parametric models: These are models that try to estimate the unknown probability distribution that produced the data. They generally require very few parameters, all of which can be calculated using a limited amount of data, and have a large amount of structure. In particular, they tend to assume normal, log-normal, or some other distributional properties. This class of models includes neural networks, regression, and time series models, but we will not focus on them here. Non-parametric regressors: These are models that are very flexible in the sense that they can seek to exactly mimic the shape of the distribution of the data and not overfit by requiring too many parameters. Trees and splines, as represented by the Generalized Additive Model, fall into this class of algorithms. Regularized models: Regularized models try to optimize the tradeoff between, on the one hand, wanting to have lots of structure in the model and, on the other, not wanting to overfit the model to the data. This is the data-driven version of the classic bias-variance tradeoff. The most common regularizing function is the L2 penalty, but the L1 penalty can also be applied, leading to a reduction in the number of active parameters. These models are attractive because they exhibit good statistical properties in learning from few data points and can also be efficiently coded from exposure headers.

3.5.2. Performance Metrics

In the context of classification models of credit or credit risk, the model is typically tested out-of-sample on another sample of loan applicants or credit risks that the bank or lender has not given credit or engaged in business with. Additionally, the sample must accurately represent the universe or population to which the model will be applied. In sum, testing out-of-sample means future applicants for credit or credit risks similar to the applicants or companies represented in the model's development sample will have characteristics and behaviors of the in-sample sample. Bureaus estimate that without

reliable out-of-sample tests, classification models for credit and credit risk have a minimal final accuracy rate of 35% to 40%.

Several performance metrics are often used to evaluate models, including the most popular measures of overall accuracy, the F-Score, and the Gini Coefficient. The accuracy of a model is the percentage of credit risk approvals for applicants with an expected default of less than 60% for a lender using a risk cutoff of 60%. The F-Score is the harmonic mean between precision and recall. Precision is the number of true positives divided by the sum of true positives and false positives. Recall is the number of true positives divided by the sum of true positives and false negatives. The Gini Coefficient is a measure of inequality which is the most popular measure of predictive quality. It measures the inequality in a population and is used to measure unevenness in income distribution. The Gini Coefficient would have a value of one if individual incomes in the entire population have no variation. In the context of predictive models, the Gini Coefficient would have a value of one if no negatively classified applicant deviated from the distribution of the labeled loan sample.

3.5.3. Cross-Validation Techniques

Several cross-validation techniques are easy to understand and put into place. They are powerful tools for finding a good model in unseen data. The most common cross-validation technique is called k-fold cross-validation. In k-fold cross-validation, we divide the available data into k roughly equally sized parts. We take one part, and we use the remaining k - 1 parts to train our model. The one part we held out is then used as a validation set to evaluate the model. We test as many samples as we have, giving us set.

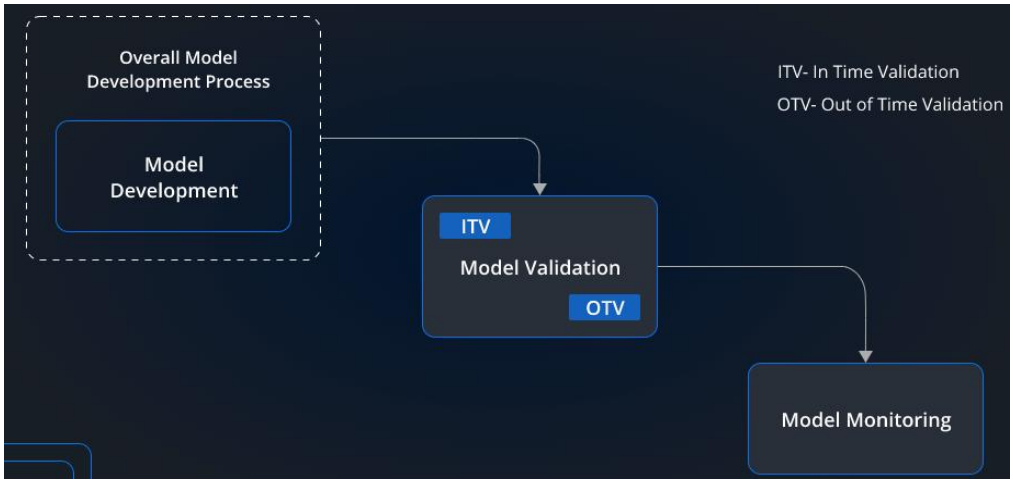

Fig 3 . 2 : Model validation in machine learning

our diagnostic set. We iterate over this process k times and at the end average over the k estimates for the model's efficacy in unseen data. This average is a good unbiased estimate of how well the model will perform in unseen data. In the case of k-fold cross-validation, the parameter k is a hyperparameter. The larger it is, the longer the validation process will take. Typically, k is between 3 and 20. Leave-one-out cross-validation is a special case. Here, the number of parts k is equal to the number of available data points. There are special types of cross-validation appropriate for time-series data often called nested cross-validation. There, what changes in the k validation process occurs in the training part of the process, ensuring that the model is never trained on the validation

3.6. Risk Assessment Models

Where ambitious, relevant, or even mandatory goals such as greater lending objectives, cost reductions for banks, or self-determined living on the part of automated learning subjects exist, they clearly justify the use of machine learning methods. This is because learning algorithms are better at processing vast amounts of data than humans. Credit applicants are of particular interest in this context. As a typical example of the consequence of machine learning recommendations, applicants have the opportunity to win an argument once they do not know why a decision is directed against them. As a welcome consequence, with regard to credit assessments, in the regulatory examination phase, customers may gain a partly self-determined life by enabling longer deviating periods, without manual refuting, or by ensuring greater distribution within bank portfolios, which has consequences for their ability to repay from the borrowers' perspective.

Despite positive aspects, there are warnings that must be taken into account. In particular, there are two aspects that must be given a different significance with regard to the design of risk assessment models, especially if deep learning algorithms are used: the prerequisites of explaining a decision or evaluation, and the adaptation to different regions or markets. Furthermore, in relation to credit assessments, the aspect of furthering preferences for certain customer groups can be seen as undesirable. In short, we highlight the positive and negative consequences of machine learning models useful for risk assessment purposes, draw conclusions about meaningful model developments such as those incorporating ethical norms, and provide some regulatory policy recommendations.

3.6.1. Understanding Risk Factors

Risk models are a way to simplify the complex interactions that influence performance. If you have identified risk factors and their relative impact, you draft loan policies to

avoid these types of borrowers and incidents. As we know, the better we can predict risk, the more robust our policies can be. Predicting risk, then, really represents the overall purpose of a credit or performance model. Since this is one of the more interesting aspects of the analytics process, this text will spend a great deal of time on this topic. But in order to fully understand how prediction works, we will have to step through a number of new concepts including a thorough discussion of variables and data that we discussed previously.

So what are these elusive risk factors that drive performance up or down? In the credit business, major risk factors are easy to understand. They include FICO score and credit history for consumer lending. FICO score is a number that summarizes a borrower's credit report and is a way lenders, insurance companies, and other entities order a credit report for uses from determining the interest rate of a loan, determining the credit limit on a credit card, or setting the premium of the car insurance policy. Additionally, other risk factors in consumer lending might include maximum years in the same job, the geographic location of the borrower, and the loan-to-value ratio of a mortgage. Each of these factors is easily understood and has a direct and material impact on how well a borrower performs.

3.6.2. Machine Learning in Risk Assessment

The assumption of purely positive relationships between predictors and outcomes is clearly the key success factor of scorecards in risk assessment. Those assumptions are sometimes violated in practice and may have significant consequences. Machine learning methods can be used for improved prediction if there are no reasons to believe in the assumptions of scorecards or if existing scorecards are performing poorly. The combination of data mining-based methods for variable generation and scorecard methods for final risk prediction allows for a more thorough model generation process. We demonstrate the results in a typical application – predicting credit risk.

A major concern in risk assessment, e.g., of credit or insurance funds, is the identification of those customers who are risky in the sense of not fully repaying their installments, or even going into default. The outcome is clearly negative, and companies and regulatory bodies pay careful attention to make a distinction between different relationships from the same negative outcome. In general, there can be various business reasons to allow not to be perfect predictors of the outcome. However, credit risk or other types of financial risks are critical determinants of regulatory requirements and related capital costs. Hence, restrictions on the allowed error rate are regularly imposed. Due to these restrictions, a large bank may not approve the credit of an entrepreneur requesting a hypothetical loan.

3.7. Implementation Challenges

Before extending the discussion to applications, it is important to understand the challenges of implementing the models that are described in this chapter. Typically, the modeler is not the decision maker for applications of the developed machine learning model. The decision makers we are referring to are the ones who have the authority, expertise, and duty to apply the developed models to rules, decisions, and policies that are used to classify and assess credit applicants. The models developed have predictive power and assess the risk or probability of default, but do not classify applicants in labels such as "approved," "rejected," "low-risk," "high-risk," or any other label necessary to make a credit decision. In many applications, human prejudice may be included—intentionally or unintentionally—into credit decisions.

There often will be bias in the target, also possibly affecting the overall modeling process. Modeling pushes one to think of decisions made by algorithms. Existing biases could inadvertently be transferred to algorithms and, therefore, to decision making. Hence, just building accurate models is not enough. Demonstrating fairness or consistently suitable performance within different relevant subgroups can be critical. Private sector algorithms may not benefit from the same public scrutiny that, for instance, criminal justice algorithms receive.

3.7.1. Data Privacy Concerns

As credit scoring moves from a credit model to a risk model, many feel a certain unease about the soon anticipated widespread utilization of machine learning algorithms. First and foremost, the confidentiality of the predictive power of more advanced algorithms is a big concern. Opponents argue that if the learned scorecard algorithm were revealed, manual credit review would reverse-engineer customers' and applicants' sensitive financial information. Adult hospital and financial benefits seekers are themselves reported to have a deep concern about having their future creditworthiness, their ability to repay, and financial vulnerability exposed to employers and providers. By law, lenders cannot use race, ethnic, national, or religious affiliations in deciding individual applicants below the materiality threshold. However, critics argue that with such powerful built-in learning devices, machine learning algorithms will unjustly incur prohibited racial, national, and religious discrimination.

3.7.2. Model Interpretability

In practice, not just making good predictions that are important, but the ability to explain how the algorithm arrived at a given prediction is also important. Lack of interpretability

of some of the most powerful predictive models is a significant limitation for the wider use of machine learning in practical settings. This is particularly true in regulated industries where the model needs to be approved by a regulator. While there is a trade-off between model accuracy and interpretability, achieving a healthy balance and making reliable predictions that regulators and decision-makers are willing to use remains critical.

In practice, we observe that most companies meet most of their model interpretability requirements through the following strategies: - If the model is already a linear model, then its explanation is the coefficient of explanatory variables. - Use simpler models. For example, decision trees, rule lists, and ordinal models. These models are much easier to understand than machine learning approaches. These models will return decisions with the appropriate associated probabilities. Moreover, if there is more than one decision rule, these models will also be able to decode the applicable rule.

3.7.3. Regulatory Compliance

Use of machine learning for credit models is gaining interest because of the potential for high predictive power. Companies need to be concerned about the reasons their existing model, which they are looking to replace, falls short. Regulators focus not only on highly predictive models but also on the explainability and transparency of the model. If a machine learning model is used to predict credit, does a predictive model meet the same standards as an existing model? What types of models are used in credit and risk prediction? Ultimately, if the model is discriminatory and predictive, will it stand up to regulatory scrutiny? The literature suggests that models that are transparent tend to be broadly acceptable to regulators and a wide range of practitioners. The most widely used models to evaluate retail credit risk are logistic regression and scorecard-based models. The outcome of a scorecard is expressed in terms of credit scores or rankings, which range from high credit quality with a low credit score to low credit quality with a high credit score. In most cases, the sigmoid function is employed to scale the scores into probabilities of default. The advantages of a scorecard over logistic regression are that it is more transparent, aligns business with risk, is easy to update, and provides security in auditing and interpretation through the use of points.

3.8. Case Studies

The goal of this paper is to show how machine learning performs when used in models that have a high level of scrutiny, review, and maturity in financial institutions. Specifically, we show the actual response models and their predictive power that are currently used by many financial companies. An overall improvement in predictive

power of 38% was achieved across a variety of performance measures commonly used by financiers. In turn, these relationships are tested against millions of applicant credit and response application records. Our results prove the actual predictive power increase of the response models used by many financial institutions that occur from the use of machine learning techniques within them. The objective of this paper is to show the impact and level of predictive power improvement that machine learning can have when incorporated directly into predictive response models, as is the case for many financial companies. In doing so, we embark upon a case study to show existing response predictive models used in many financial situations and how improvements have been realized where models have gone live. In addition, sample base criteria for response metrics will then be used to provide readers with the results and predictive power improvement achieved after incorporating machine learning into the models.

3.8.1. Successful Implementations

Having covered a broad spectrum of topics in this book, we can now present some implementations of machine learning techniques suitable for credit risk models. We use real-life examples of successful implementations using seven scoring techniques. The performance of these techniques is measured using data from credit card application problems. Data from a direct marketing campaign is used to illustrate expected profits. The main goal is to detect first which model performs best within the same dataset. We use a proper gain chart to find the performance within each dataset. The information gain is expressed in a concise way such that everybody gets a vivid and transparent perception. The experimental results of the credit risk problem indicate that tree ensembles, such as gradient boosting, dominate neural networks or high-dimensional data on this dataset. We also provide a discussion of when to use which methods and the potential of combining different methods for subpopulations.

3.8.2. Lessons Learned from Failures

The economic consequences of flawed credit and risk models could be severe, not only individually but also because they can result in systemic banking failures. The financial crisis is an example of the potential consequences of fast mortgage lending growth, declining underwriting standards, growing housing price bubbles, and the rapid expansion of mortgage-backed securities. Foreclosure losses reached significant levels, exceeding individual mortgage lender equity levels. Many diverse mortgage lenders, financial institutions, and investment banks would either fail, merge, or require assistance. The problem is that predictive models can perform well in sample or designed test selections and fail out of sample, for example, in different cross-time, cross-

sectional, or macroeconomic environments. Prediction evidence would suggest that certain evaluations failed to consider relevant risk mitigation and economic indicators.

There are many possible reasons why lenders fail to appreciate the implications of risk models that often are based on history and not on future judgment. For example, except in the severe housing bust that contributed to the financial crisis, assuming home prices will always increase is reasonable. Risk models may be technical and not easily understood, especially when they fail to provide a broadly accepted simple solution that blames problems on external factors. Regulations can allow financial institutions to sidestep capital equity requirements. Financial institutions also have promoters, accounting, and rating agencies as potential allies. Regulatory and policy standards may grow too lax to allow the rapid overall market and housing price declines that require a business credit recession to fully correct. Policymakers do not consider how many mortgage lenders and homebuyers would need assistance that exceeds reasonable foreclosure expectations, differentiated market prices, and financial support. Policymakers might focus on the primary goals of promoting widespread affordable homeownership and the potential benefits from a rising economy. Regardless, public decision makers should anticipate potential problems with any overbought macro sector. Banks with underfunded deposit insurance that are too big to fail require immediate government emergency financial support due to a potentially systemic loss in confidence. These factors and others foster a favorable lending environment that can allow credit risks to grow rapidly until they fail.

3.9. Future Trends in Machine Learning

Now that machine learning is firmly ensconced in a favorable light, where is it going in financial services and beyond? What are some intriguing new capabilities? How may new initiatives develop? These questions are food for pondering, especially given the rapid pace not only of machine learning techniques but also of real computing power to apply them. Here is what I think may develop in the next few years. One major area deserving broad attention is the use of novel data sources. Of course, there already are many initiatives recycling traditional data streams in new ways, and these can have significant predictive power. But it's important to look at what new sources of data may emerge as the digital age progresses. Technology, generally. With the profusion of new risk models being exploited, a hotbed of new potential modeling areas is the need to integrate real, measured operational risk. Moreover, most operational risk systems will need to forecast operational outcomes – over several months, not just using the imminent lens of crisis modeling, i.e., the logistical risks leading to a hit to the value at risk.

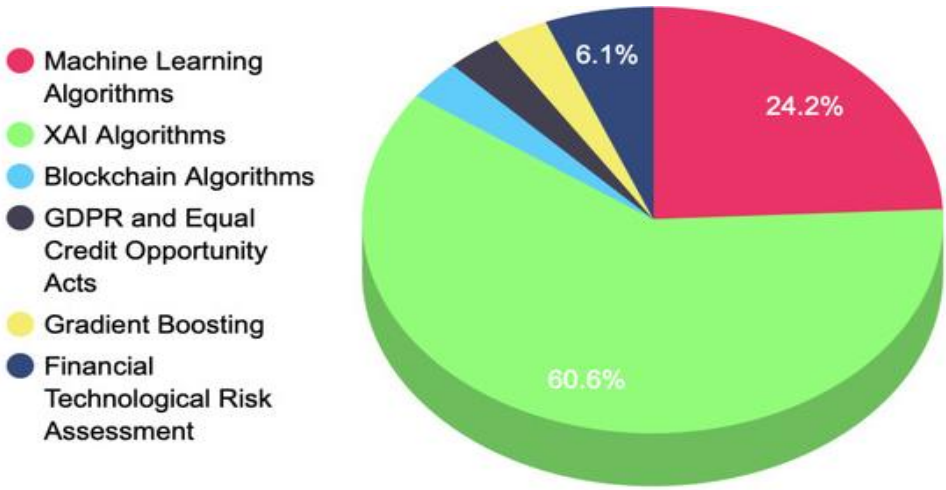

Fig 3.3 : Credit Risk Assessment and Financial Decision Support

3.9.1. Emerging Technologies

Nine years ago, as the CDO and the big data analytics leader at Citigroup, I started a very early exploration of the big data technology stack, particularly the Hadoop open-source stack. Together with some colleagues and a few vendors, we built strong momentum in which the entire data analytics community at the time sat up and took notice, and some are still taking notice. In the following years, many other companies embarked on similar efforts and created a market. That market has shown steady growth in order of magnitudes. The statistics can be utterly boring at times due to quantity and repetition. The market for big data enabling tools and technologies clearly falls under this description. What makes these statistics and prognostications fire up the imagination is the inescapable trend that the technology stack underlying the big data market is being embedded in all the other software businesses of the time. This will happen faster than you think.

Those embedded applications will find their big data technologies being complemented by machine learning and predictive modeling that serendipitously utilize the big data infrastructure for hyper performance. It is no exaggeration to say that the business of predicting what will happen is now the hot business, forcing a convergence between predictive modeling libraries and big data clouds. Predictive modeling business solutions no longer evolve like moving glaciers; they wink into existence when conditions become optimal. You will find in this chapter an overview of machine learning and predictive modeling technologies that are of interest to stakeholders in or have an interest in the big data market. These stakeholders will find examples that demonstrate the kinds of science that have to be done to make these technologies deliver value to their businesses.

3.9.2. Predictions for the Credit Industry

Machine learning is an effective tool in predicting outcomes related to financial services applications, either as a tool to improve binning efficiency or in generating new attributes. We demonstrate this by using the credit scoring file to predict the loan terms for new applications. Credit scoring is the tool used by banks to quantify how much they trust a loan applicant. The bank will have a policy of lending to applicants above a given score but not to applicants below it. These scores are generated using attrition data on previous applicants. In our case, we note the outcome score where the applicant gets credit. Using att2 now as the target, we attempt to predict loan outcome results and to suggest the major attributes that could potentially be used instead of or in addition to the outcome score. In their credit scoring file, we see that the scorecard attributes of "Previous Tel" and "St" (most common previous occupation) are both important and valid scores for the risky duration. We further identified a number of additional attributes that we do not see on the scorecard list and that could potentially be important in predicting the risk duration, including "St" and "LikeTel".

3.10. Conclusion

Predictive power is a key metric used across the risk management industry, and the ability to generate accurate measures of credit risk is a key competitive differentiator for these institutions. Maximizing the accuracy of credit models is the job of many workers, not only in risk management, but also in the fields of machine learning, artificial intelligence and statistics. With the advent of high-performance computing, and frameworks and algorithms, predictive modeling techniques provide an additional layer of accuracy and insight into credit and risk models already constructed by methods.

In this paper, we have introduced advanced quantitative techniques that are turning many heads in the risk management space and presented a unified approach at blending traditional risk management tools and newer machine learning algorithms to both solve for predictive power and provide understanding, transparency and insight into the sources of that predictive power. We have demonstrated on multiple data sets that ensemble models are also able to generate surprisingly high accuracy levels, a sine-qua-non for choice algorithms in a fast-moving world, where several challenges need to be addressed if a large proportion of our banks are to be compliant with regulations, and thereby economically competitive over the next few years. We are creating Business As Usual workload and improving our credit and risk performance for our institutions as part of our new Basel reality. Are you?

3.10.1. Final Thoughts and Future Directions

Our main conclusion is that machine learning on retail credit scoring datasets does indeed have predictive power, and different models have strengths and weaknesses based on the actual default rate. We found a suite of algorithms that is generally better at classification than logistic regression and is particularly better with heavier and more imbalanced datasets. However, while citing performance alone is likely to precede finding models that work best for their particular application, attention also has to be paid to heteroskedasticity and interpretability.

Several avenues remain to be explored. Firstly, data preprocessing. So far, we have employed only a few simple strategies such as capping and winsorizing and rescaling for a few variables to account for differences in the underlying distribution of the training and application populations. Different variable transformations or transformation grids without any effect would improve the quality of the program. There are also other score transformations that can be explored, although their impact is much less with an average centered at 600 as applied in our problem.

References

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). *Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research.* European Journal of Operational Research, 247(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Bastani, H., Ascarza, E., & Choudhury, P. (2021). *Predictive Modeling and Machine Learning in Credit Risk Management.* Management Science, 67(11), 6785–6806. https://doi.org/10.1287/mnsc.2021.4103

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). *Consumer Credit-Risk Models via Machine-Learning Algorithms.* Journal of Banking & Finance, 34(11), 2767–2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

Sirignano, J., & Cont, R. (2019). *Universal Features of Price Formation in Financial Markets: Perspectives from Deep Learning.* Quantitative Finance, 19(9), 1449–1459. https://doi.org/10.1080/14697688.2019.1571683

Malekipirbazari, M., & Aksakalli, V. (2015). *Risk Assessment in Social Lending via Random Forests.* Expert Systems with Applications, 42(10), 4621–4631. https://doi.org/10.1016/j.eswa.2015.01.002