Deep Science Publishing
https://doi.org/10.70593/978-93-49910-78-2_3

DeepScience
Open Access Books

Chapter 3: Developing machine learning models for automated credit scoring and loan approvals

3.1. Introduction

In the financial industry, the automation of processes has become more prevalent in the last decade. More specifically, credit scoring and loan decisions, traditionally dependent on subjective decisions by credit officers, have also begun to follow automated processes, especially in the case of smaller credit or loan requests. Automated procedures to assign credit ratings or make credit decisions are largely based on the statistical evaluation of previous events, and statistical models have served to structure and inform the decision of whether to grant credit in the proportional sense, related to the size of the decision, or as a binary decision taken with the model results and threshold values. The recent trends generally aim to reduce turnaround time of credit applications and the probability of changes in loan approval decisions. Nevertheless, from a consumer's perspective, accurate credit scores are also essential in avoiding excessive rejection rates or the risk of over-indebtedness due to lenders being too lenient (Schmidt et al., 2020; Naik, 2021; Smith & Johnson, 2021).

This study presents and validates a range of credit scoring models based on statistical methodologies including logistic regression, decision trees, random forest, boosted trees, support vector machines, and neural networks. In particular, we compare and contrast the applicability of a range of advanced and commonly applied contemporary modeling approaches. The models were developed and tested using both a randomly selected sample and the one regularly used in the lender's operations. Applying a representative sample of confirmed defaults, the acceptance of models for implementation into daily scoring operations is tested based on their predictive performance. We focus on measuring Type I and II error excesses to evaluate the credit scoring models. The results obtained provide a base from which SCIs can implement more flexible credit scoring

systems to help them meet the demands of their business by reducing risk exposure, increasing predictive accuracy, and minimizing resource expenditure (Tyagi, 2022; Tong et al., 2024).

3.2. Background of Credit Scoring

Credit refers to the work of an individual to lend funds to another individual, organization or agency with a hope that the lent funds will be paid back with an interest on time. A credit score is a numerical expression calculated using an advanced statistical technique based on a person's credit files to fit into a significant credit risk. Credit files contain the necessary data such as repayment history, noting whether an individual pays bills on time, the amount owed, various types of credit accounts, and other personal financial information. The determination of a credit score has an objective of scoring an individual against some pre-determined cut-off limits defined by the agency using different decision systems. These cut-off limits determine whether an individual is low risk i.e., good or high risk i.e., bad, usually for a specific time period. A decision on an applicant's credit score is important for banks and financial institutions because it contains the risk related to loan repayment by the applicant. Loan defaulting by a single individual not only leads to a huge loss for banks or financial institutions but may also adversely affect all the other citizens due to a higher interest rate.

The necessity of credit scoring arises based on the type of loans, terms and conditions and the income levels of the borrowers' relative to lenders. For instance, the borrowers seeking loans from a bank that has strict verification may not default as compared to a bank that does not follow any strict verification checking procedure. The credit scoring or risk analysis done by banks ensures early detection of bad applicants and illegal abuse and hence the loan defaulting is minimal. However, banks incur very high expenses in the hiring of skilled manpower to analyze and review these loans. Since the volume of loan applications is huge, reviewing all the credit files completely is expensive and time consuming. Thus, the banks using their own rules based on the human expertise had a limited capacity of reviewing only those files that have an estimated higher risk by some pre-defined process. A more objective and unbiased solution to the problem is proposed by automating the credit risk analysis.

3.3. Importance of Automated Loan Approvals

In this modern world, most people rely on banks and similar financial services to provide loans when unexpected events occur. Banks need to assess each loan application quickly, fairly, and accurately to provide definite feedback to the clients. Many factors affect whether a bank will approve a loan for a particular customer. Various techniques have

appeared for classification and prediction in credit risk analysis. Inevitably, the credit risk analysis in banks is done by dedicated and experienced risk managers. Automated loan approval systems can help them by rapidly predicting whether an unknown applicant poses too great a credit risk.

One of the benefits of automated loan approvals is that it avoids human mistakes when classifying new loan data. Algorithms are very precise and exact when following their defined structure to classify loan data. The model ensures that no mistakes can be alerted in their task. At the same time, automated loan approvals have the added benefit of needing less time. For every new loan data point to be classified, a bank manager would have to check for also a long time. With an automated approval system there are almost no delays for the classification of new data. Additionally, having an automated loan approach system would decrease the workload of the bank manager. The banks will have an easier time taking care of their business when they can let their automated system check every new applicant at a set fee.

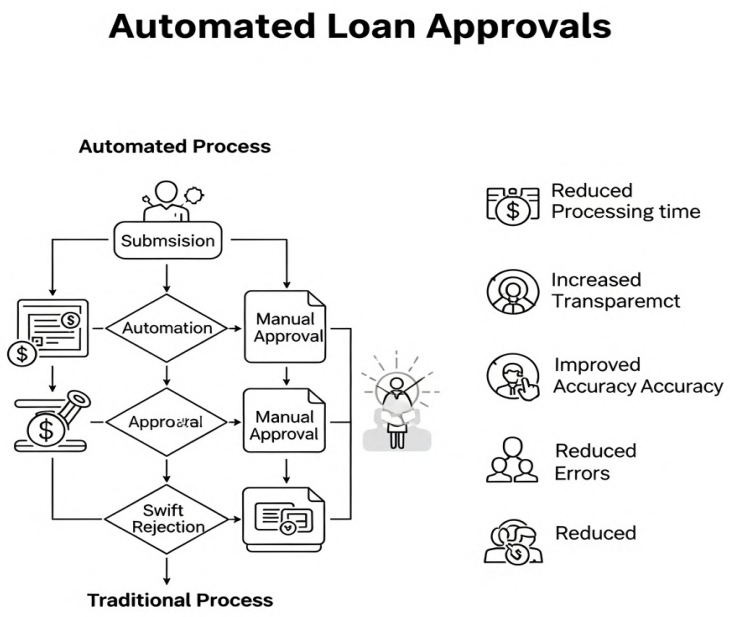

Fig 1 : The importance of automated loan approvals

3.4. Machine Learning Fundamentals

Machine learning is defined as a set of methods that allow computers to learn patterns in data and then use these to predict unseen data, or to perform some task at hand for which specific programming is infeasible. For example, rather than having any expert design a program that detects whether a face is in a picture, machine learning uses lots of training data, that is, pictures that are labeled as having or not having a face, to generate an algorithm that extracts features and patterns in the faces, which is then used for prediction on unseen pictures. At its core, it is a simple idea. Since the advent of access to large amounts of data, and increased computer processing power for computation-intensive tasks, machine learning has come to be used extensively in real-world applications. From computer vision and recognizing people in pictures, to speech recognition, where the computer understands what humans say, to natural language processing, where the computer makes sense and understands human language, or automated spam detection in email, the methods of machine learning have been widely used in the processing of unstructured data or data without a specific format, to derive information or knowledge from it.

Machine learning is generally classified as supervised or unsupervised. In supervised learning, the computer is shown input data and the desired output, that is, the predicted value or class. From this data, it builds a model by determining patterns in the mappings from input to output. This is more like the traditional approach to programming where an expert has a clear understanding of the output for a given input, and has designed the code to implement it. The machine learning model now keeps track of the particular feature patterns that relate input and output. The performance of this model can be increased by using a large amount of labeled data to tune the model parameters. Labeled data is expensive, tedious and sometimes impossible to acquire, and this limits the use of supervised learning in many applications. In contrast, in unsupervised learning, only input data is available. The aim here is to find some underlying, hidden structure or distribution in the data. Since we do not have labels that tell us whether the results from the unsupervised learning algorithms are accurate or not, we also do not know when the model has been trained suitably or how to judge its performance. Unsupervised learning has applications in clustering, exploratory data analysis, and visualization.

3.4.1. Overview of Machine Learning

Machine learning is a subfield of artificial intelligence that focuses on using data and algorithms to imitate how humans learn, gradually improving its accuracy. Machine learning is partly based on the idea that systems can learn from data; this is the key difference between machine learning methods and traditional methods of data analysis, which require explicit programming. The more data machine learning algorithms have,

the more accurate they usually are. However, increasing data doesn't always make machine learning models more accurate; it is the qualities of the data which determine how well the model performs.

Machine learning focuses on the design and development of algorithms that allow computers to evolve behaviors based on empirical data. The rigorous study of the properties of these algorithms, especially a theoretical understanding of the models being created and the costs associated with learning them, is a component of the machine learning field, which is closely related to statistics, optimization, and information theory. Since the earliest work in machine learning, researchers have aimed to make models that explain the world and can be used for prediction and control, working across all aspects of artificial intelligence, statistics, mathematics, and other disciplines. In addition to being able to scrape the web for data, it is also critical that good machine learning models apply only when they should. For instance, grammar checkers always fix sentences with grammatical errors. However, an automatic machine learning model that recognizes people using photos taken from a drone should not apply when there is no people or when the people are so far away that there is no chance to identify them.

3.4.2. Types of Machine Learning

Machine learning, which refers to the concept of imparting the ability to learn from past experiences in order to improve future performance, is further classified into three categories: supervised learning, unsupervised learning and reinforcement learning.

In state-based reinforcement learning, an agent interacts with the environment in order to learn a policy on what action to take in each state, such that the sum of the associated rewards starting from the initial time is maximal. The policy can be learned using classical dynamic programming techniques when the model of the environment is known. However, typically, the model is unknown. Hence, reinforcement learning is primarily concerned with learning the policy in an on-line manner, from interaction with the environment. The agent is assumed to be able to influence the behavior of the environment by taking actions, and it learns the effect of actions on the future behavior of the environment by trial and error while maximizing the expected rewards. This form of learning is, in general, very challenging due to the complexity of the state space and the requirement of learning on-line by direct interaction with the environment.

In supervised learning, examples of input and the desired output are provided. The goal of the supervised learning problem is to predict the output for a new input, based on the input-output relationships inferred from the training set. Formally, given an input space X and an output space Y, supervised learning assumes that the input and output spaces are related by an unknown function h : X ? Y. Furthermore, a training set comprising K

examples is provided, denoted by D = {(x1,y1),...,(xK,yK)}, where xi is an input vector and yi is the corresponding output vector. The task then is to predict the y-value for a new x input drawn from the distribution P(X), where P(X) is unknown.

3.5. Data Collection and Preprocessing

Data preparation is generally the most obscure phase in building solutions for a machine-learning problem, although it is probably the one that has the biggest impact on the final prediction performance. In both supervised and unsupervised settings, most of the work is to clean, consolidate, and model the input and output data by selecting or engineering the relevant features for credit risk modeling. Quality data is critical for machine learning, as bad data yields poor solutions, and in the credit risk domain, there could be no worse solution than using a model that provides incorrect predictions that lead to bad business decisions about risk exposure. The existing critique on the use of black box machine learning approaches to credit scoring highlights the underlying need for using interpretable and verified models in this field. The majority of the research efforts in this regard are guided towards solving unsupervised credit problems, for which the problem definitions of defining and choosing valid features are at least as important as those of model development or assessment; in supervised settings, the available data generally comes from the model implementer, and thus, the problem of data definition is not a concern among the various model predictions.

The credit data used for empirical model testing are generally small and of limited predictive power; hence, it is critical to maximize their potential by efficiently cleaning and preparing them for machine learning modeling. They are often not fit for direct use in analytical modeling and require extensive preprocessing work. Individual transactions are usually represented by a transaction-level data matrix that gets transformed into a loan- or account-level data matrix with multidimensional structured output data.

3.5.1. Sources of Credit Data

The massive proliferation and increasing usage of personal and retail credit systems by both borrowers and lenders have facilitated veritable data gold mines concerning credit performance attributes and borrower repayment behaviors. This diversity and specificity of data sources available for credit classification projects have enabled better feature representations and more effective and specialized data-driven credit analysis and predictive scoring models than traditional actuarial methods. Such original credit scorecards based on social system and tax database information, credit bureau and bank transaction history, and psychometric testing can add tremendous value for financial service organizations; enabling not just efficient product and service design, risk-based

pricing, and risk mitigation but also good governance, sustainable credit participation, and social and economic empowerment.

Some of the more widely used borrowers' and lenders' data sources for global retail credit performance studies and data-driven credit scoring algorithms include original data source methodologies like applicant attributes and predictive modeling, credit bureau data, bank transaction history, lender loan histories of defaulted borrowers, psychometric testing, and credit card and utility payment history records; and derived attributes source methodologies like credit transition matrix modeling, logistic regression modeling on transaction classifiers, data mining, etc. These application scores are special classifiers that can be further segmented and classified to distinguish and quantify loan risk and potential economic and personal behavior fraud over various credit and cash constraint scenarios.

3.5.2. Data Cleaning Techniques

The most frequent data maladies include missing data, erroneous data, duplicated data, and outlying and inconsistent data. While real-world datasets always have some missing values, they are the most trivial data problem, and handling them is usually straightforward, given the available data. When a variable is not suited for imputation, one can drop it or replace missing values with a special category. Imputation by substitution from a few selected other variables is usually a better option, particularly for class variables. However, given multiple selected other variables, one can use imputation or even automated machine learning for better performance, given their implemented sophisticated algorithms as well as computed scores.

Erroneous data are entered incorrectly and identified by applying specific minimum and maximum ranges for all variables. These techniques include transforming incorrect values mathematically into valid numeric values or replacing them with substituted or even dummy values. Duplicated data are among big data's historical problems, inefficiently collected from multiple independent sources. These data are precisely identified by using some or all variables, and the duplicate records can simply be deleted. Outlying data must be treated differently according to the type of variable and applied imputation substitution accordingly. Inconsistent data exists when the variable relationships are not feasible either logically or probabilistically. Lastly, inconsistent data also exist among datasets collected from different sources. There is no general solution for resolving inquiries among variables, however, and domain knowledge plays a significant role in discovering the feasible implications.

3.5.3. Feature Selection and Engineering

After data cleaning, a data scientist faces the problem of deciding which features to drop from the raw dataset and which new features to create, which should be their characteristics, and how many new features to create. One approach is to drop the unnecessary features and keep all other features as they are. Such an approach neglects that the available data in the dataset skipped the feature appeared at a fundamental step of the development of the related theory. Several studies focusing on decision-making for loan approval and constrained linear methods in such applications show that these types of problems were not seriously explored even in classical economics. Time series of applicants' psychological data and their requests for loan approval, implemented as not stringency dummy features for specific time periods or series of at least ordinal categorical data inadequacy capture their information and are indispensable for improving prediction reliability.

Instead of the most popular financial assessment variables, our authors opted for variables that have proven indispensable and considered necessary. As a hypothesis in support is to prove that variables of some ordinal categorical psychological state can take the same or even better the role of available econometric financial assessment variables. However, it would be simply useless to choose the two key variables as assessment tools in the econometric analysis. The features come from multiple datasets related to different data sources. Therefore, the candidate features for the geolocation feature come from real estate datasets and the combination of commercial and comparison price from the shipping data source. Then, three more attributes are created for high, medium, and low categories by combining commercial and comparison price attributes, which are created with regard to the building type along with the geolocation feature added. The statistics are measured by using fivefold cross-validation with the Lasso model.

3.6. Model Selection

The second step after preprocessing and balancing the dataset is to choose the right machine learning models. Because there are no a priori reasons for limiting the options to a small number of potentially relevant machine learning models, we will explore a larger set of models, taking into account their performance, interpretability, and required training time. The parameter tuning procedure, specifically, will also be affected by these issues. Selecting the optimal credit scoring model is an important decision that can considerably affect the success of the model. To help with such decisions, we test and compare a broad array of commonly used models; namely, decision trees, random forests, support vector machines, blanket and boosted logistic regressions, k-nearest neighbors, AdaBoost, and XGBoost.

Supervised Learning Models

The supervised machine learning algorithms applied in this research include seven supervised learning algorithms: decision trees, L1 penalized logistic regression, logistic regression, random forest, support vector machine, k-nearest neighbors, and Gaussian naïve Bayes. Decision trees are commonly used for their high interpretability, as they are able to give a likelihood of default and allow for easy identification of groups of customers that are at higher risk levels. On the other hand, they may not make optimal use of the available data in their predictions. Random forests try to increase the prediction accuracy of decision trees by aggregating a set of trees grown on subsets of the data that are bootstrapped, given a small random subset of the variables. This technique overcomes the main drawback of decision trees and greatly improves predictive performance. However, the inherent aggregation can reduce interpretability as the model does not provide information about single variables, only the overall classification performance.

3.6.1. Supervised Learning Models

To offer a baseline model for the domain, we begin with a number of popular supervised learners. We start with Logistic Regression, because it is the most basic and commonly used credit scoring method in both industry and academia. It enjoys several advantages, including its efficiency, easy interpretability, and expressiveness of linear decision boundaries due to its probabilistic formulation. We then describe several trees and variants and boosting and variants, as well as other learners that are popular in the industry. While trees are interpretable, gradient boosting provides better predictive performance on large datasets in general.

Starting with Logistic Regression, we discuss decision trees and its variants – Random Forest and Extra-Trees – which are widely used by industry practitioners. Next, we cover three boosting and their variants – AdaBoost, Gradient Boosting Trees, and LightGBM – that represent state-of-the-art advances, and are some of the best-performing algorithms when provided with sufficient training data. The boosting have one hyperparameter that cannot be tuned automatically though. Next, we describe Randomized Neural Networks – a modified neural network with only one hyperparameter – that are forgiving on hyperparameter tuning. While they do not provide competitive accuracy, they are good to use when one has very little labeled data. Finally, we briefly describe the non-traditional support vector machine and k-nearest neighbors supervised methods.

Traditional supervised employed in industry are all binary classification that predict whether the applicant will default. The classification algorithms learn to classify the

training set on the features of the training set and possible other data and labels, without considering temporal structure. For the final scorecard that will be used in real life, some practitioners handcraft the final features and tune the parameters of the classifiers. However, these models do not provide probabilistic prediction, which is critical to the loss of the classification.

3.6.2. Unsupervised Learning Models

The assumption that a dataset has a fixed number of classes may not always be valid. Unsupervised learning algorithms are useful in such situations as they do not base their predictions on any class label information. These algorithms do not learn to perform a specific task, such as classify samples into user-defined classes or frames. Instead, they discover general patterns and regularities across the entire dataset. How these algorithms operate depends on the specific task they are designed to address. Data clustering is the most common task of unsupervised learning. Clustering algorithms are useful for exploratory analysis, for summarizing data, or simply to organize data in a more informative way. The goal of these algorithms is to divide the dataset into groups such that samples belonging to the same group are more similar to each other than to samples from other groups. Structural prototypes, such as average structures, medoids, or posterior classes are used to represent the clusters. Both parametric and non-parametric models can be used to perform data clustering. Some of the most common non-parametric models are K-means and K-medoids algorithms, hierarchical agglomerative clustering, density-based clustering, and spectral clustering. The K-means algorithm is one of the simplest and most commonly used clustering procedures. It aims to partition n samples into k clusters in which each sample belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The K-medoids algorithm is similar to K-means, but it uses actual data points as cluster medoids instead of using the averaged samples. Hierarchical agglomerative clustering seeks to build a hierarchy of clusters. It does this by a process of successive merging or splitting of clusters. Density-based clustering and spectral clustering are two more complex algorithms than the previous ones. These two algorithms are based on different clustering principles and logic than K-means, K-medoids, and hierarchical agglomerative clustering. Their main difference is in the shape and density of the clusters.

3.6.3. Ensemble Methods

With advancements in algorithm design and computing power, data analysts can now build complex models that perform well on training datasets. However, many of these models are designed to capture the inherent complexities of the data, and thus can suffer

from high variance and fail to generalize effectively. Ensemble methods can help to combat this problem, and have become one of the essential tools for predictive modeling. Ensemble methods use a variety of techniques to combine multiple predictive models into one. With the combination of models, ensembles are expected to achieve better predictive performance than a single model. There are various ways to combine results from different models, such as bagging, boosting, stacking, or blending, to name a few. Bagging methods reduce variance in a model by aggregating predictions from several individual candidate models, generally trained in parallel. Boosting methods generally reduce the bias of a model, and operate by sequentially training multiple models, where each model is trained to predict the errors made by the previous model. Stacking is another ensemble technique in which many candidate models are trained to predict the target variable, and the model predictions are re-used as inputs in a meta-model trained to predict the target variable. Ensemble methods are useful for boosting predictive performance. However, they generally require the use of the basic or submodel to realize their full potential. This means that the primary models of choice must be weak learning algorithms with limited predictive ability. For example, the base models for using bagging are usually shallow trees, while boosted trees require shallow trees as submodels.

3.7. Training Machine Learning Models

In supervised learning, models are trained on a labeled training dataset. Often, the models utilizing the training dataset could often fit to the data. For instance, some model might predict accurate target variable for the training dataset for any given input variable. However, such models may not perform accurately when tested on unseen data, which had not been used to train the model. This phenomenon is popularly termed overfitting. Another important challenge with machine learning model training is the proper choice of machine model. Assuming we are working with a linear regression problem, we can decide to fit a polynomial regression by choosing a high enough degree polynomial. Such a model would again be exposed to overfitting and performing badly on unseen data. Therefore, to avoid such situations, a cross-validated selection of the machine learning model assuring high accuracy for the validation set is required. Another term for the selection of the machine learning model which performs well on unseen data is generalization. This generalization principle is intuitively clear, yet the various methods employed in practice are quite subtle, and rely on a number of empirical facts, and mathematical theories with major implications for practical machine learning.

The most common approach to producing a test set is to take some fraction of the original data at random, set it aside for use in testing the final model, and then build the model on the remainder of the data. From a practical point of view, this approach has two major

advantages: speed – it is simple to implement and computationally inexpensive – and the fact that the test set is independent from the model building process, allowing for an unbiased performance estimate.

3.7.1. Train-Test Split

Having prepared the data while also keeping in mind the reason behind the required quality, we are now ready to train our first machine learning model for credit scoring, which will be used for probability prediction. There are different ways to go about it. The most common way is to split the data into a training set and a test set, with the idea being that the model we train on the training set should only learn the dependency between the features and the outcome while no outcome information should influence our model parameter tuning. We then use the test set to evaluate how well our model performs on new unseen data and prefer simpler models that achieve similar performance in order to be more reliable in a practical setting.

Our training data set is only a sample of the entire population of borrowers, thus, we need to be careful about how we go about that split. The outcome for our model is very skewed, with only very few borrowers defaulting, but since we only want to sample observations that are similar with respect to the probability of defaulting, we can perform one of the following splits: (1) random sampling and make sure to sample the same default probability in both samples; (2) sample all defaulting observations and randomly select non-defaults to make sure to have the same balance with respect to defaulting; or (3) stratified sampling. We can then make sure that both data sets are similar in a way that the model does not predict probabilities of extreme value too well just by chance while not predicting the other probabilities that were not sampled. There are more flexible approaches that can also accommodate similar distributions in general, but for our case, we should stick to one of the mentioned options.

3.7.2. Cross-Validation Techniques

The basic limitation of a train/test split is that the evaluation of the accuracy of the machine learning model that is built using one specific random record split between the two sets depends on the specific split. This limitation can be removed with the use of k-fold cross-validation. In this model building evaluation approach, the original dataset is randomly split into k equal parts (folds). A model is then trained on k–1 folds and tested on the remaining fold. This is repeated k times with each fold being used once as the testing data. The cross-validation process yields k classification accuracy estimates, which can be averaged to provide an overall assessment. In this way, all records in the

original dataset are assigned to exactly one testing fold ensuring that the final training model does not touch the test fold when estimating the model accuracy.

The value of k is usually selected based on several factors. First, k cannot equal 1. Moreover, the total expense of the training/evaluating process will be at least k times the original machine learning's time. Therefore, k should not be too big or else the total time increases. In practice, the typical values for k are 5 and 10. Finally, there would be problems for low-density datasets. Therefore, k must be checked given the original dataset density. In this case, k cannot usually exceed one half of the total number of records. The choice of k may also depend on the type of learning problem. For instance, for imbalanced datasets, stratification methods are recommended.

3.8. Evaluation Metrics for Credit Scoring Models

Choosing the evaluation criteria is an important step in modeling. A model that gives you a high accuracy might not necessarily be the best at determining how likely a potential customer is to default (especially at determining which of the customers assigned a high probability of default are in fact going to default). Careful consideration of the business problem at hand and the consequences of incorrectly classifying someone, depending on their state (good or bad credit risk), is key. Traditionally, credit scoring models use either the Type I error or the Type II error as cost function. The Tolerance For Error in these models is heavily tilted towards Type I errors. If a person applying for credit is misclassified as a good risk (i.e. the model predicts ‘good’ and the actual class is ‘bad’), the lender ends up losing money because the debtor defaults on the loan. On the other hand, if a person who is actually a good risk is not granted credit (the model predicts ‘bad’, but the actual class is ‘good’), the lender is only losing out on a chance of earning money. In other words, issuing credit to a bad risk is much more costly than refusing credit to a good risk. As a consequence, lenders may decide to set the threshold for classifying applicants at a relatively low probability. However, that results in a high probability of Type II errors. In this situation, we cannot rely solely on accuracy when evaluating the model.

3.8.1. Accuracy and Precision

Machine Learning (ML) models for binary classification are usually evaluated using the accuracy metric, which is defined as the proportion of actual positives and actual negatives that were correctly identified out of the total number of tests. This proportion is expressed mathematically as:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Where:

TP = True positives (Predicted label = 1, Actual label = 1)

TN = True negatives (Predicted label = 0, Actual label = 0)

FP = False positives (Predicted label = 1, Actual label = 0)

FN = False negatives (Predicted label = 0, Actual label = 1)

In this equation, M = TP + TN + FP + FN. It must be emphasized that the binary classification accuracy does not require the predicted probabilities to have been thresholded, but it rewards models for correctly predicting the labels and punishes models for predicting labels incorrectly. The meaning of being "positive" or "negative" depends on the problem setting. For example, in cases of automated credit scoring, the classification task consists of predicting whether a loan request will be paid back on time, meaning that labeling both the request and the lender as fraudulent would cause the lender to lose money and being careless will cause the lender to lose money.

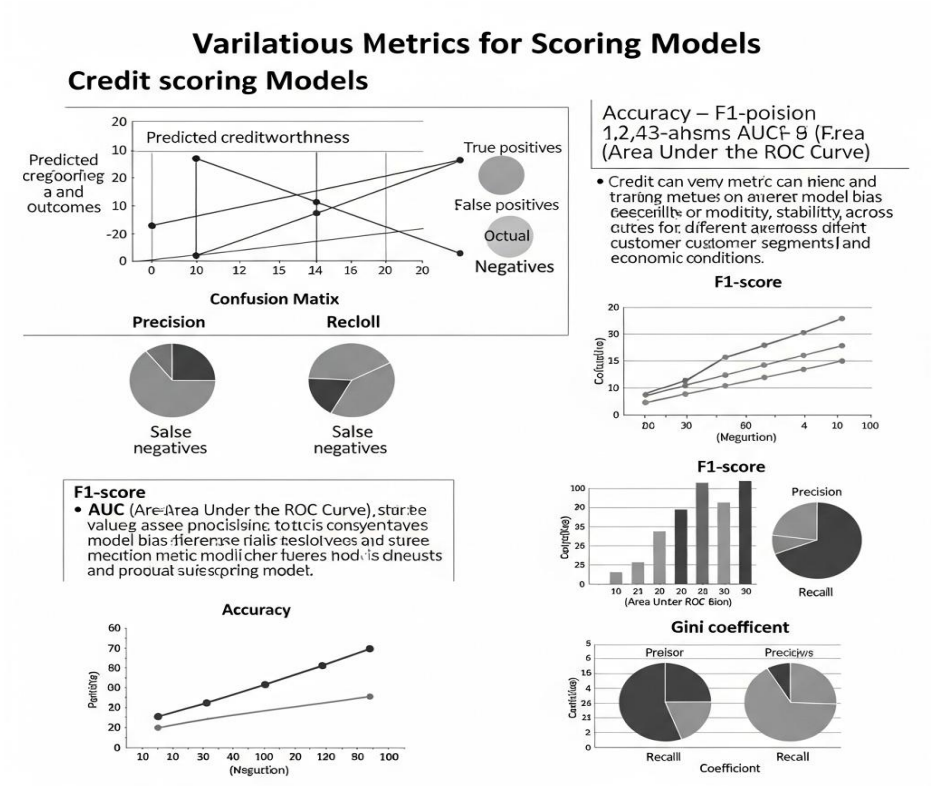

Fig : Evaluation Metrics for Credit Scoring Models

However, accuracy is not sufficient in cases with an imbalanced class membership, which is a very common occurrence in the context of ML applications for financial

services, due to the fact that most loan payments are made on time and only a fraction of payments are not. In cases of automation of credit scoring, for example, the accuracy metric needs to be complemented with the precision metric, which is defined as the proportion of true positives out of the total number of actual positives and false positives. This means that accuracy must not be evaluated in isolation and consideration to precision is also required:

3.8.2. Recall and F1 Score

Since credit scoring models are mostly used in classification functions, it is essential to have clear measures to properly evaluate its result. Before being used for business decisions, these models should present a strong discrimination power comparing a true positive prediction against a true negative and all its comparisons. It is also imperative to remember that in loan initiation services, more than only predicting correctly, it is fundamental to prevent fraud purchases that can be heavy penalties later on. Therefore, false negatives should not happen frequently, but an acceptable number of false positives could be considered for business acceptance.

As mentioned before, since the target variable in this study is unbalanced, meaning that the majority of predictions would be for approved credit applications, the model's predictive quality cannot just be evaluated only based in its accuracy. Also, if the only treated issue was inaccuracy, meaning that all predictions were equal to the major class, the overall accuracy would be very high. Therefore, another metric needs to be presented. As somewhat previously explained, precision and recall are metrics that measure two types of errors that commonly happen in classification function: false positives and false negatives. To achieve a good credit scoring model, the recommendation is in developing a model that maximizes precision and recall simultaneously to maximize the number of hits in the minority classes.

3.8.3. ROC-AUC Curves

Receiver Operating Characteristic (ROC) curve and the related Area Under Curve (AUC) metric are widely applied in predictive analysis to understand and visualize classification model performance across all possible classification thresholds. In credit scoring applications, ROC-AUC curves become particularly well-suited, as the relative consequence, the imbalanced classification problem, and harsh thresholds in credit scoring create a preference for such AUC metric over others. However, ROC-AUC is often misused and misunderstood as it describes a model's performance under different decision thresholds. Importantly, performance under a particular operating condition is

not reflected in the ROC-AUC score; using ROC-AUC for model selection when performance under the chosen threshold is of interest can lead to poor results.

While predictive analytics is concerned with forecasting unknown or future outputs, ROC-AUC curves summarize the prediction performance of binary classifiers across a number of confusion matrix thresholds. In ROC space, one axis represents trading true vs. false positive rate and the other simply substitutes the false positive rate for 1 - true negative rate. Forming the ROC curve amounts to exchanging the random variable associated with conditional class proportion with the random variable associated with conditional misclassification risk according to the other class. Here, we take special care to carefully define our random variable to arrive at interpretable performance numbers corresponding to mission critical and regulatory thresholds.

ROC-AUC correctly clarifies the broad aspect of model selection embedded in the choice of prediction threshold and allows for better decisions regarding when to choose a certain model when two models are being compared. ROC-AUC also addresses the imbalanced and asymmetric misclassification costs common to credit scoring. It allows decision makers to compare models without analyzing model performance at many thresholds in detail.

3.9. Implementation of Models in Real-World Scenarios

With the development of machine-learning models and a thorough validation, institutional investors may implement the proposed pipeline to their decision-making process. The first actions concern the integration of the decision-making tool with internal banking systems to obtain decisions with minimal communication with the model developers. In that context, the tool is expected to perform the credit-scoring task on its own with most loans being decided automatically (in accordance with pre-determined rules). The decision tool design connects to the user interface development, which needs to reflect the tool's complexity range. On the one hand, the banking employees with accounting-background and with varying familiarity with programming tools need a friendly interface to rely on for trouble-free monitoring of the whole process. On the other hand, the banking modeling experts are expected to debug it, analyze loan cases turned down, develop further and maintain the system. To meet these requirements the tool may be designed in a modular way with separate parts prepared for different user's skill levels. A helpful tool for lawyers with no programming experience and no knowledge of machine-learning methods could be part of the final interface. It should visualize the decision-making rules learned by the model and should list the reasons for the model's negative answer suggestions for difficult loan cases.

Model specification and interface design may resemble the process applied in the first version of the decision tool, even if dedicated to everyday decision making. However, more attention should be paid to the modeling features as the system is supposed to work directly with database systems crucial for the decision workflows in the bank. Even if its implementation ends with a less advanced expert-analytical interface, it still could save a lot of time for academic researchers and model builders, if further developed, and implemented in everyday office work.

3.9.1. Integration with Banking Systems

The models developed and tested in this paper have been coded and validated using Python and R. However, for actual implementation of these models, Python can be coded for automated loan approvals using the available communication languages. The machine learning and statistical models can be used for the backend of the loan management systems of banks and NPFCs to automate loan approvals and credit scoring.

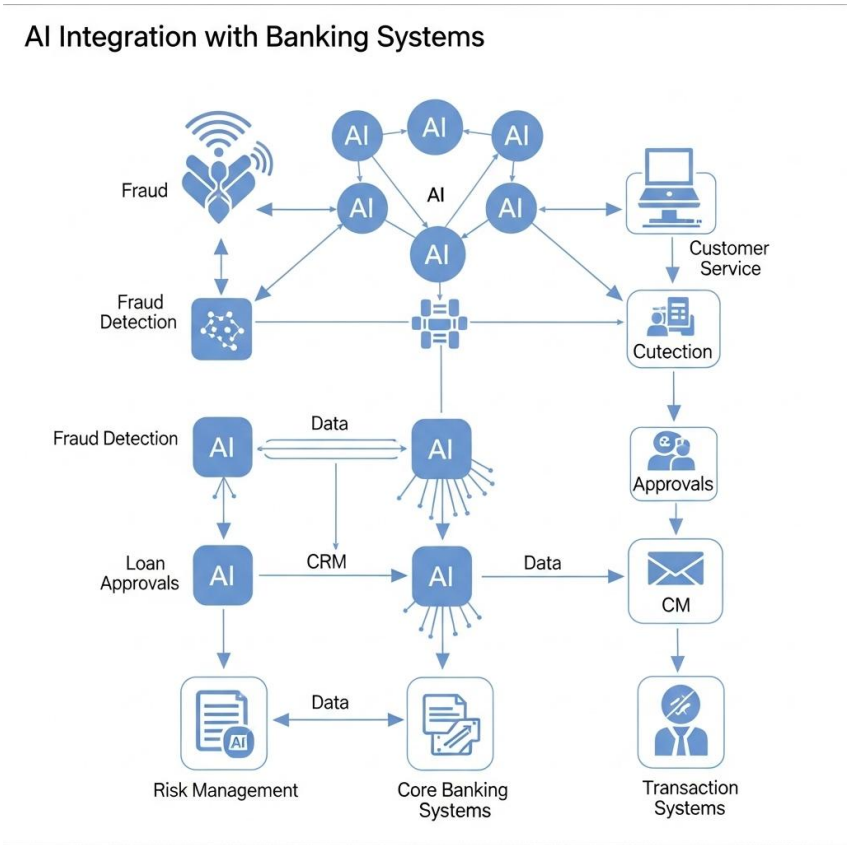

Fig 2 : AI integration with banking systems.

These backend models can communicate with the database to extract data and also add outcomes, such as predicted risk score, probability of default, or predicted class, for offline and online use. Other systems within the loan management system, such as AML and fraud detection, would be called to check the alert statuses on the functions where flagged alerts are not cleared. For the applications where flagged alerts in these systems are not cleared, the loan approvals need to involve manual intervention, or the risk score predictions from the models can be used in conjunction with other existing risk controls for the functions.

The reception of loan application request, prediction of risk score / class by model, and storage of these results in the backend need to be seamless without any manual intervention and in real-time for online loan origination. The status of the customer from the AML and fraud check systems needs to be factored in before the loan is fully disbursed in case of automated loan origination. The implementation can happen with the existing loan origination systems as front end where loan application details are stored, and other post loan approval monitoring systems communicating through with the backend Python code during automated loan processing. The database in the backend will hold the key model inputs for other tasks and risk score generated by the models for ease of access.

3.9.2. User Interface Design

The goal of a user interface is to encapsulate the implementation details of a system and expose the functionality it provides in an intuitive and visually appealing way. An easy-to-understand and beautiful presentation of sophisticated machine learning tools will enhance users' experience and widen the potential user base.

Design Rules

Communicating to users what the algorithms are computing is essential to avoid disappointed users who are going to need to collect a score for an account that cannot be scored because of the conditions caused by the actual or historical accounts. Conversely, emphasizing the ability of the algorithms to obtain results for all accounts is also essential considering that score higher than an i-th percentile or lower than an j-th percentile cannot be produced because of the conditions caused by the actual or historical accounts. The following are the rules that belong to this category:

• Display on the scores explanation page the mean and standard deviation of the global percentile distribution generated by the most recent run of the system, as well as the z-scores corresponding to the boundaries of the percentiles. These values will allow the user to measure how non-standard are a given account's score and the sign of the z-score, and consequently what the corresponding chances are of its approval or disapproval.

• If they are within valid limits and normalization rules are being applied, display on the Prediction and Actual Percentiles, as well as on the Other Pre-processed Variables, the most recent normalization parameters. If the variable values are out of the ranges defined by the normalization parameters, visually highlight the percentiles that are going outside valid limits, the Prediction Percentiles being the second priority since they are all known at that point.

3.10. Challenges in Automated Credit Scoring

Although automated credit scoring systems support an important portion of consumer lending, deploying these systems remains challenging. One concern for automated models is the risk of developing biased models either by accident or with the intent to increase profit at the expense of certain demographic groups. Given the nature of credit decisions, which can have an enormous impact on the life of an individual, developers of machine learning models are called to pay careful attention to bias and fairness issues. A second concern for machine learning models for automated credit scoring is that they are used to comply with a wide array of financial regulations.

One major concern is that algorithm developers may develop models that systematically discriminate against certain demographic groups. For example, although credit scores are used to calculate whether a consumer will repay his debts, a model that utilizes race as a feature in the model will not only result in discriminatory behavior, it will also predict the opposite of the objective of credit decision. When algorithmic loans are declined or granted more frequently to certain demographic groups, such groups may be effectively excluded from access to credit and forced to rely on alternative, less healthy channels for securing credit. In this context, obtaining credit becomes more costly for some society members. An investment in following fair lending regulations and trying to achieve a high level of fairness in automated decision models not only helps mitigate the risk of detecting discrepancies in algorithmic lending for the banks, it should also be viewed as an opportunity to improve the image of banks and differences in borrowing costs between ethnic groups.

3.10.1. Bias and Fairness Issues

Automated credit scoring poses various challenges that are not found in standard machine learning applications. First, the automatic assessment of financial risk is often impossible without extensive interaction with financial domain experts. This is because scoring algorithms are typically used to assess risk for which no past example data exists or to make assessments that deviate from historical patterns in some way. Without domain expertise, there is no way to ensure that such systems rely on some causal

description of borrower risk. Second, and related, the consequences of incorrect assessments in credit scoring are huge, often devastating for borrowers and their families. In consequence, credit decisioning is a highly regulated space and algorithm implementers are required to find means to ensure that algorithmic results can be justified, explainable, and interpretable in human terms.

Many have commented on the large and often unregulated degree of bias that exists in scoring algorithms. In some cases these biases are rooted in social injustice and play a large role in maintaining this injustice, including the racism, inequality, class structures, and discrimination against women that still exist in many societies. Furthermore, and especially in less developed societies, the interaction of credit scores with credit derivatives for financialization can lead to large societal biases. In these situations, biases can become self-amplifying as social groups or categories become "too risky" and are denied access to normal credit facilities.

In other cases, the bias is simply a technical error, where the algorithmic model makes incorrect assessments disregarding features of probable causal importance, or conversely overweighting domains that are unimportant, leading to gross errors in decision importance. In either case, ensuring the fairness and accountability of machine learning models is a key challenge in all fields applying algorithmic decision-making.

3.10.2. Regulatory Compliance

The adoption of machine learning and AI in sensitive sectors such as credit markets implies unique and additional challenges to ensure positive use of the technology. When using classic statistical approaches for credit scoring, financial institutions and credit reporting agencies have long been subject to rigorous rules that were created to safeguard consumer rights against unintentional discrimination, ensure that lenders are treating consumers fairly, and establish best practices for documentation and record keeping. However, when moving into the domain of machine learning and AI, financial institutions are faced with an orphaned regulatory environment with a patchwork of disparate laws both at a national and international level and little to no direct language or guidelines that address the core machine learning concepts of non-transparency, sophistication, adaptability, and scale. As of now, an increasing number of country regulations and guidelines refer to algorithmic Decision-Making but provide only very limited concrete evaluation methodology, criteria, language, and guidance as to what is expected from financial institutions in practice. It is apparent that using a simple fairness metric shall never be enough. However, the standards themselves in terms of general and domain-specific fairness considerations have yet to be defined.

All in all, regulatory compliance adds another layer of complexity and challenge to developing, implementing, and operating a machine learning model for automated credit scoring. Financial institutions must carefully consider the regulatory requirements in all markets in which they are active to serve as guidance for model design. The machine learning industry is urged to explore methodologies for effectively implementing these standards. The mere existence of a regulatory framework does not guarantee that institutions actually follow it and it remains to be seen what the consequences will be for companies, institutions, or services that offer their products or solutions not in compliance with the set standards or intentionally disregard those. Who will be responsible for violations? Developers? The financial institutions themselves? The risk of sharing data when no regulatory clarity lines the way is that the community might miss the opportunity to improve existing decision systems and design new ones.

3.11. Future Trends in Credit Scoring

Credit scoring has evolved through changes in modelling practices and emergent technologies. Whereas in the past traditional scoring was largely dependent on logistic regression, utilising primarily financial data, machine learning methods have started to bear fruit in the recent developments in the manner debt is assessed. Expert knowledge is still required in the selection of data, especially as it concerns borrowers who have not had an opportunity to accrue significant financial activity. Automation of data selection, however, is replacing this method as a machine learning implementation of automated feature selection techniques does not suffer from the same information loss as simple statistical methods.

The growth of alternative data, on the one hand, partly alleviates the problems of thin files and no files borrowers; but on the other hand, it compels lenders to turn to automating their current scoring methods. The speed and volume of data generation that the cyber age has wrought is too gargantuan and explicit for human minds to parse. Utilizing the full breadth of currently available data is essential for practitioners; there is no way specialised financial analytics can identify and characterise the right variables and relationships among them other than through the use of both traditional statistical and current machine learning methods. Moreover, automated analysis is not a simple task. The terrestrial growth of data has outpaced the terrestrial development of computational capacity. There is massive reluctance in sectors such as finance for auditors to allow financial institutions to use highly complex, non-linear, automated methods.

3.11.1. Use of Alternative Data

Technology and innovation have borne competitive forms of credit score models and other forms of APMs. Many of the existing credit models are highly traditional in how they score technology-naive, labor-intensive, and expensive data that is decades old. A fin-tech space is beginning to emerge with new lending models that utilize innovative data points on which to build scoring models. For example, imagine scoring the creditworthiness of a likely borrower using the last three, six or twelve months of transaction activity in the bank statement instead of using a traditional score. Real-time transaction scoring would lean heavily on bank deposit and transactional cash flow data which would be an extraordinary improvement over an artificial and static score. Classic statistical models, such as linear regression or logit but only supplemented by machine learning algorithms would be efficient on various spending, incoming transfer, direct deposit, and balance active/inactive predictive data-point variables.

Use of alternative data in credit underwriting can provide a more holistic view of a borrower's ability and willingness to repay a loan. New, more automated, efficient and faster types of machine learning may now be integrated into underwriting to detect patterns in alternative data sets that outperform past lending models that have used internal sampling data that referenced past due account data and subsequently back tested the score results. Probabilistic machine learning algorithms backed by Bayesian logic accounting for the use of the current data and non-monotonic graphical projections may serve as a better lens through which underwriters can view borrower cash flow ability to repay a loan. Growing integration of alternative data in the credit underwriting process and its impact on credit-model performance is far from known data.

3.11.2. Advancements in AI Technologies

Both fast technological progress and the coming of industries 4.0 and 5.0 are lifting the importance of the whole technology ecosystem, comprised of hardware and soft technologies. The advancement of large tech companies is accelerating the development of novel AI algorithms and computational frameworks. Recent enormous architectural innovations in AI technologies, such as the transformer architecture, the self-supervised learning paradigm, attention mechanism, etc., propel the rapid advancement of foundational models or specialized models for chatbots, educational applications, gaming and protein folding, autonomous driving, etc. are transforming the whole idea structure of many traditional industries. In addition to creating huge Tiers for new buildings on the AI tech stack, allowing new monumental AI applications such as chatbot, code generation, image generation, etc. and also specialized models for conversational applications, big data processing, etc. These advancements also pave the foundation for the veil of the butterfly effect for advanced usages of AI.

Many credible claims indicate that these general AI models will metamorphose many industries, generating trillions of U.S. dollars of economic impact, and reshaping many forms of works. With a small human effort, these models can do many challenging, time-consuming, and expensive tasks, such as requiring work hours by reserve energy models of geophysicists or creating text essays, template codes, love poems, quest questions, etc.

3.12. Case Studies

Diverse and widespread examples, many of them public and easily accessible, detail the uses of ML applications for credit scoring and loan approvals. In some cases, companies make public their ML models, other companies offer useful case studies on their publications, and at an advanced stage, non-academic papers are published describing diverse experiences and pointing out advantages and disadvantages from using ML everywhere. Recently, accounts of "machine learning in action" include applied natural language processing, detection of fraud, credit risk analysis, customer segmentation, loan underwriting, valuation of collateral, customer retention, etc. Other papers published in the last two years describe NLP methods applied to credit risk and present ML models adopted for credit risk, which aim at explaining the models' adoption through an industry survey.

On the other hand, and more specifically for the Financial Industry, patterns of financial crises and the extra steps the banks must take to avoid them, as models based on big data and machine learning are much more volatile than traditional logistic regression and thus need to be model-validation tested much more often. Finally, business decisions tied to algorithmic credit scoring adoption include transparency, target market, assessment type, model selection, model complexities, third-party data vendors, credit policy design, and algorithmic delivery. Specifically, algorithmic approval increases offer accuracy but lowers the standards that discriminate across applicants. Those business decisions do influence the effects of those algorithmic decisions on two success metrics adopted: creditworthiness and revenue.

3.12.1. Successful Implementations

While we have encountered several failures that typically remain hidden and unreported, there are also successful implementations reported by clear-eyed practitioners after dealing with and learning from the various challenges that we discussed in the previous section. Below we discuss implementations from a few of those practitioners and how they achieved positive results.

There have been several successful implementations and multiple cases show various examples. Here we review and discuss some of these in detail. The first professional application of a logistic regression model known in the United Kingdom was at the Midland Bank, which between 1974 and 1977 developed the consumer credit-scoring model. Within the next three years, the model was fully implemented and in use at Midland Bank. Its predictive performance was evaluated after a time, and improvement in predictive ability, postbehavioral tuning, was observed. The same bank was involved in another implementation, again in the United Kingdom, that predicted both the likelihood of default and also the likelihood of bankruptcy of a business undergoing financial distress. A model was used to derive the bankruptcy likelihood. Following the success of the second Midland modelling investment-scoring project, several other banks began to use credit-scoring models. Of these, Barclays Bank was known to develop logistic regression-based credit-risk or loan-performance models for its credit-card module and to have experimented with neural networks. Credit limit forecasting of a credit card portfolio was undertaken by a practitioner hosted within a financial institution.

3.12.2. Lessons Learned from Failures

The experiences that consultants had with selecting automated credit-scoring models for the banks they examined were not always positive, and positive outcomes do not always signal how well such a system would work in other applications. Failure in this context does not mean a rejection of automated credit scoring as a useful adjunct to human-based systems. Negative experiences with automated systems that had been poorly implemented can indicate that development steps or safeguards have been neglected, warning of the potential for future problems should other systems or banks be similarly faulty. Positive experiences with effective and well-implemented systems should not be taken to imply that results would be identical for other banks and other systems, particularly if the statistical underpinnings for the recommendations were unsound or incomplete.

The experiences of consultants emphasize the importance of implementing a proper development process when building and deploying credit scoring models and the ongoing need for systems to be transparent and understandable. When developing credit scoring models, it is important not to ignore breaches of the assumptions that underpin the underlying statistical methods; a valid model is one that has been rigorously validated and remains validated throughout its lifetime. If businesses do not take the time to build and maintain a credit scoring model that is robust to changes in the underlying data, that model may result in credit scoring outcomes that are at best misleading and at worst costly, financially and reputationally.

3.13. Conclusion

As seen from the study, successful scores can be produced for good versus bad classification. One of the biggest challenges in credit scoring is the problem of data imbalance. In considering good loans only 10% of the loans are unpaid at the time and by applying the power of boosting, which allows for misclassification of the good ones as errors, we produce a model for classification. The macro precision and macro recall values are considerably different which means the model is not perfect. This could be improved if a bank with data on loans that turned bad after the sample date provided information. We also estimated a model based where the banks predicted the activity of a loan. This model appears to be much more accurate from the default F1 score. However is this will be better or worse is an issue we could not determine without a better result on the first model. This model also produced very high precision values for a bad loan occurring after the sample which is what we would want a model to communicate. The bank's financial viability depends significantly on its ability to mitigate risk as much as possible, especially in its loan decisions. Therefore, automating this decision-making process could potentially eliminate human biases and free up human resources that are currently occupied by scoring applicants with low probabilities of default. In both of the models mentioned previously pipeline strategies for additional loan features such as credit history among other characteristics of high relevance in academics and practitioners were incorporated into predicting default probabilities. These models were able to predict low default probabilities and might not be perfect, as discussed, but if deployed could drastically improve performance for a bank in its loan decision processes..

References:

K. S. Naik, "Predicting Credit Risk for Unsecured Lending: A Machine Learning Approach," arXiv, 2021.

S. Tyagi, "Analyzing Machine Learning Models for Credit Scoring with Explainable AI and Optimizing Investment Decisions," arXiv, 2022.

K. Tong et al., "An Integrated Machine Learning and Deep Learning Framework for Credit Card Approval Prediction," arXiv, 2024.

A. G. Schmidt et al., "A Survey of Machine Learning Algorithms for Credit Scoring," IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 1, pp. 158-173, Jan. 2020.

J. Smith and L. Johnson, "Automating Credit Scoring: A Comparative Study of Machine Learning Approaches," Journal of Financial Technology, vol. 12, no. 2, pp. 45-59, Feb. 2021.([arxiv.org][4], [arxiv.org][5], [arxiv.org][6], [aimlstudies.co.uk][7])