Methodological Notes on Scorecard at origination Development

Bilel Ammouri

August 16, 2023

Contents

1 Introduction 2

2 Scorecard WORKFLOW 2
2.1 Raw Data Collection . . . 2
2.2 Data Preprocessing . . . 2
2.3 Data transformation . . . 2
2.3.1 Weight of Evidence . . . 4
2.3.2 Feature selection . . . 4
2.4 Sampling . . . 5
2.5 Training Model . . . 5
2.5.1 Logistic Regression . . . 6
2.5.2 Cut-off . . . 6
2.6 Scorecard . . . 7

1 Introduction

The "ScoreCard," like other predictive models, is a tool used to assess the level of risk associated with applicants or clients. While it doesn't identify applications as "good" or "bad," it provides "Odds" scores or probabilities that an applicant with a given score will be "good" or "bad." These probabilities are then used as a basis for decision-making.
There are three key statistical approaches in the "scorecard" for retail portfolios, which are of paramount importance in the application of the new IFRS 9 standards (aimed at predicting the behavior of existing customers regarding the payment of their future installments based on past payments and anticipating any future default by creating provisions). These approaches are as follows:

- **Score at Application or origination**: This is the most important statistical approach in the "scorecard" for retail portfolios. Its objective is to assign a score to the customer when granting credit, reflecting the risk of default for that customer. The significance of this score lies in its ability to enable the lending institution to accept or reject customers when granting credit based on this calculated score.
- **Behavior scorecard**: Another statistical approach in retail credit scoring, in this case, the behavior of customers to whom credit has already been extended is studied (reassessing the credit risk of the customer, taking into account recent behavior in addition to the variables considered during credit origination). Compared to the origination score, the behavior score is dynamic in the sense that it changes over time based on behavior variables and any data that can enhance this score.
- **Dynamic scorecard**: For the dynamic score, credit risk is evaluated at any future point in time, unlike the origination score and the behavior score, where risk is evaluated at a fixed time horizon. This type of score requires more advanced statistical techniques such as survival analysis (these techniques allow us to not only determine if a customer will default but also when).

2 Scorecard WORKFLOW

The project steps will be classified as follows (see Figure 1):

1. Raw Data Collection
2. Data Preprocessing
3. Data Transformation
4. Sampling
5. Training model
6. Cross Validation
7. Final model

2.1 Raw Data Collection

This project draws upon the valuable resources of open-source data from both Kaggle and the UCI Machine Learning Repository, augmenting our capability to construct a robust origination scorecard.

2.2 Data Preprocessing

Data preprocessing plays a pivotal role in machine learning projects. In the context of the "scorecard at origination" project, the data is of a cross-sectional nature, leading to potential anomalies that encompass:

i missing values,

ii outliers,

iii Multicollinearity.

Various techniques exist for managing missing values, including options like deletion, imputation, and modeling. In order to determine the most suitable approach, a test is performed subsequent to correction, or a comparison is made between the densities prior to and following the correction process (see figure 2).

2.3 Data transformation

Data transformation constitutes a foundational stride in the preparation of data for analysis or machine learning. Among the many techniques available for data transformation, binning is a prevalent approach. Binning involves the segmentation of a continuous variable into distinct bins or intervals. This technique proves particularly advantageous when the aim is to convert continuous data into categorical data or to streamline data for analytical purposes.
Benefits of Binning:

i **Simplification**: Binning serves to streamline intricate relationships by converting continuous data points into discrete categories, thereby rendering them more amenable to interpretation.

ii **Noise Reduction**: Binning aids in diminishing the influence of noise within the data by aggregating similar values, resulting in a more consolidated representation.

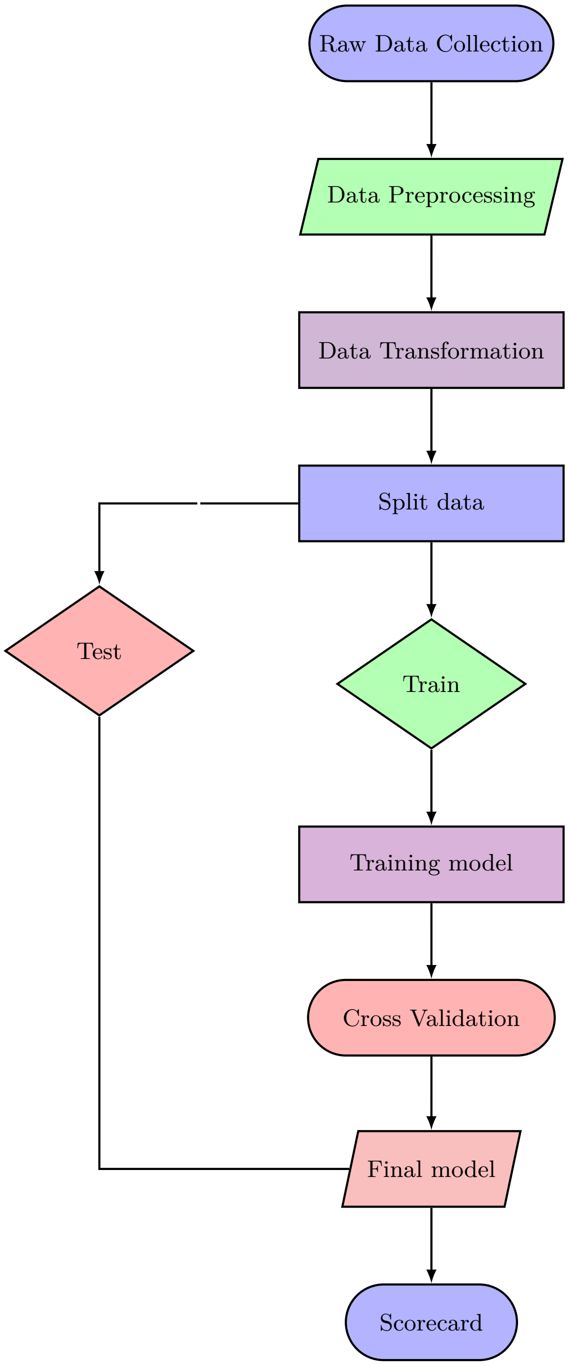

Figure 1: Workflow process

iii **Categorical Handling**: Binned data can be treated akin to categorical data, permitting the utilization of techniques specifically designed for categorical analysis.

iv **Visual Depiction**: Binned data lends itself more readily to visualization through mediums such as histograms or bar charts.

Benefits of Binning:

- Label Encoding/ Ordinal Encoding
- One Hot Encoding
- Dummy Encoding
- Drawbacks of OHE and DE
- Effect Encoding / Deviation E / Sum E
- Hash Encoder
- Binary Encoding
- Base N Encoding
- Target Encoding

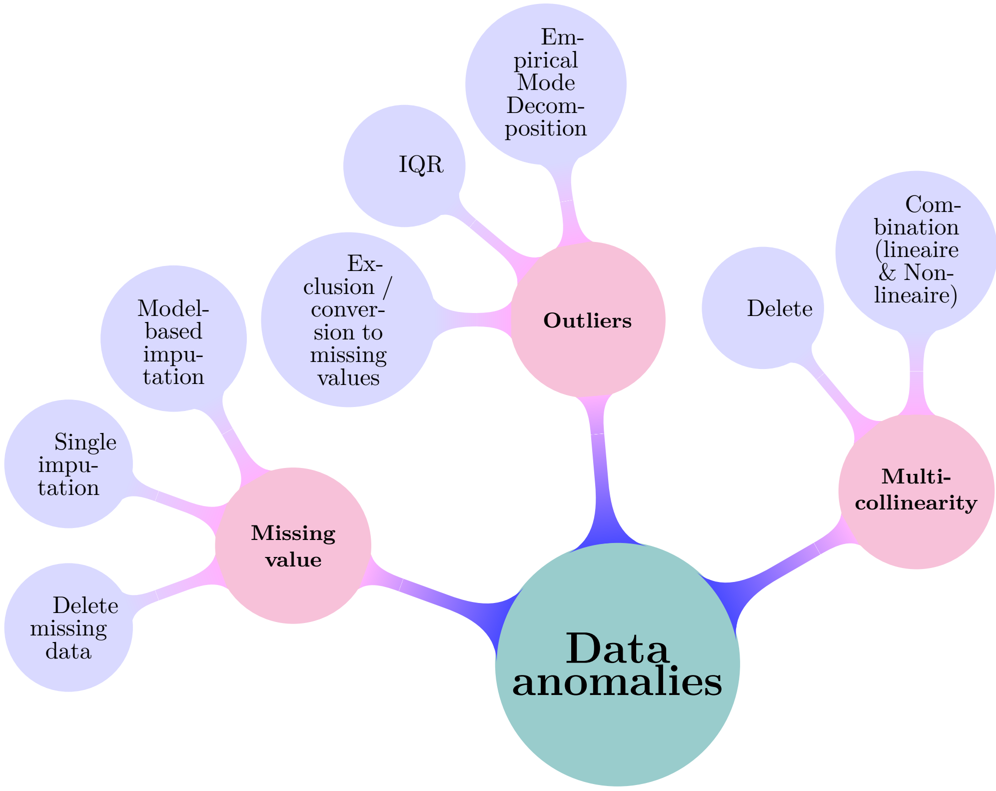

Figure 2: Data anomalies correction

2.3.1 Weight of Evidence

Note

The Weight of Evidence (WoE) indicates the predictive strength of an independent variable with respect to the dependent variable. As it has originated from the credit scoring domain, it is often described as a metric reflecting the differentiation between positive and negative outcomes. The terminology of "positive" and "negative" outcomes might not be readily applicable in the credit risk context. To enhance clarity, we can expound on the WoE concept using the framework of events and non-events.

Formula

$$WoE = \ln\left(\frac{n_g(i)}{n_b(i)} \times \frac{N_g}{N_b}\right) \tag{1}$$

With,
N_g : the total number of "good guys" in the population;
N_b : the total number of "bad" people in the population;
$n_g(i)$: the number of "good" in category i;
$n_b(i)$: the total number of "bad" in category i.

2.3.2 Feature selection

Numerous techniques exist for feature selection, with many of them being referenced to facilitate a performance comparison. The ultimate algorithm to be embraced will incorporate the variables selected during this phase.

Note

We provide the loan officer with the flexibility to choose the variables.)

This project involves supervised classification, with a predefined number of classes (2 classes), encompassing both quantitative and qualitative variables (conventional). The selection of the feature selection algorithm will hinge on its compatibility with classification models, considering aspects like filters, wrappers, and embedded methodologies (see Figure 3).

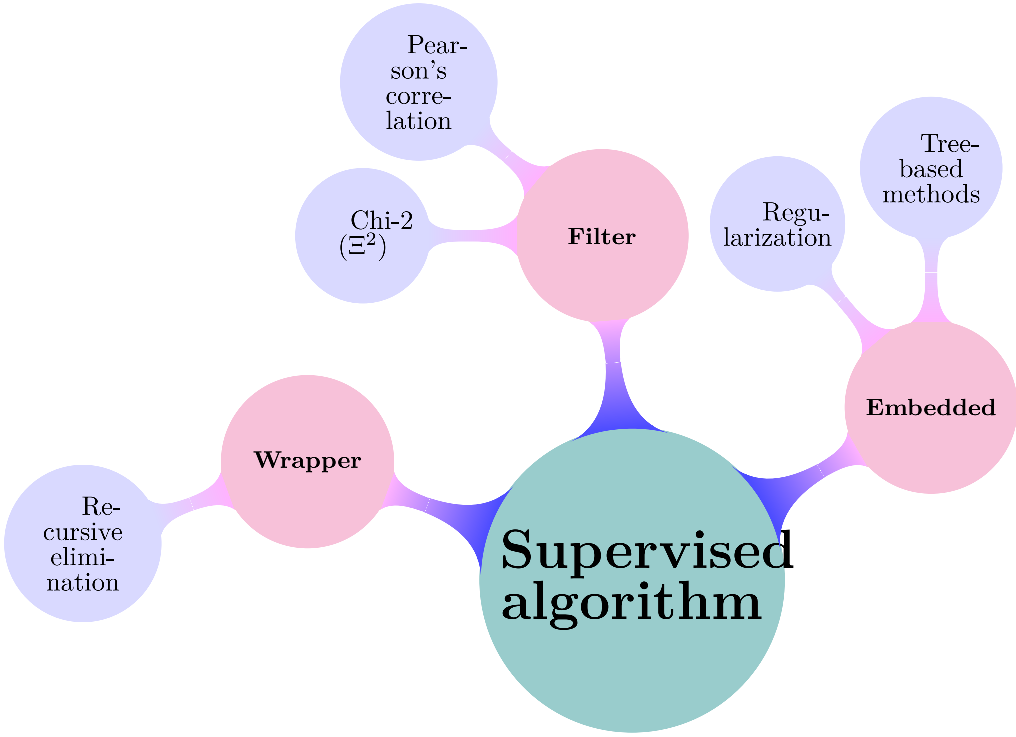

Figure 3: Feature selection

2.4 Sampling

To substantiate the chosen technique/model, validation must be conducted utilizing a sampling approach. This entails partitioning the sample into two distinct sub-samples: the learning sample and the test sample. Typically, these sub-samples are allocated 3/4 and 1/4 of the original sample, respectively.
This project classifies the credit applicant into two classes (rejected/accepted). Couple of cases present either a balanced database (both classes have the same weight, which isn't generally the case) or an unbalanced database.
There are several « sampling » techniques classified between « under » and « over » sampling (see Figure 4).

Note

To assess the robustness of the selected algorithm, both balanced data (utilizing various sampling techniques) and unbalanced data are employed for analysis.

2.5 Training Model

Numerous alternative methods have been employed by researchers in addressing classification challenges within credit management. These classification approaches can also be utilized to determine the optimal separation between the two classes, namely 'good' and 'bad' (binary credit approval decision). Notable examples include graphical and longitudinal models, 'bump hunting' techniques, Markov chains, nearest neighbor clustering algorithms, probit and tobit models, genetic algorithms, linear programming, genetic programming, and support vector machines. However, there is limited evidence to suggest that any of these methods significantly outperforms traditional approaches when applied in practical scenarios, or that they are widely adopted within the credit industry.
The classification algorithms employed are as follows:

i Logistic Regression
ii Decision Tree
iii Neural Networks: MLP Classifier
iv KNeighbors Classifier
v Discriminant Analysis
vi CatBoost
vii Bagging Classifier
viii Gradient Boosting Classifier
ix Random Forest
x AdaBoost Classifier
xi Gaussian Naive Bayes
xii SGD Classifier

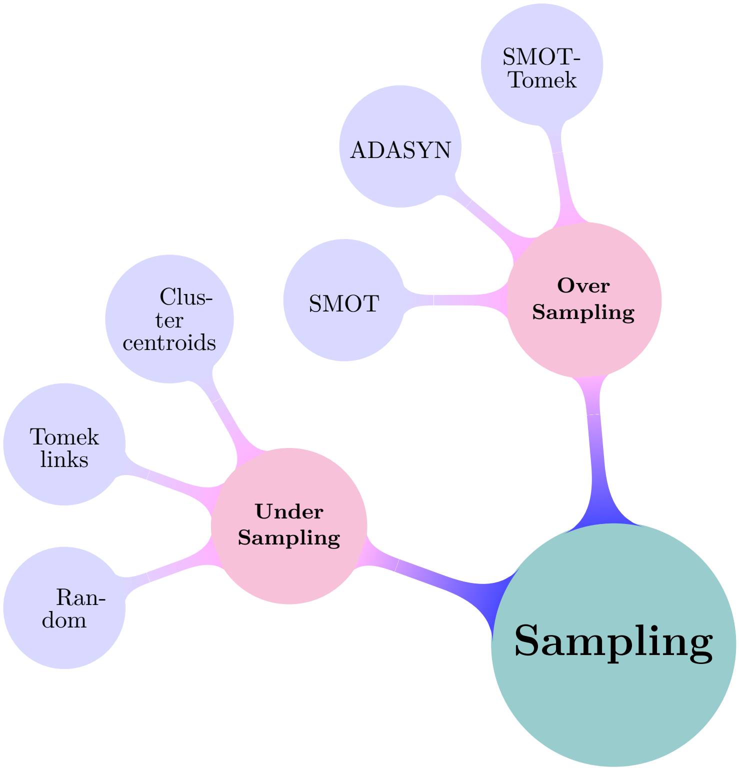

Figure 4: Sampling Techniques

2.5.1 Logistic Regression

Note

The utilization of logistical regression in scorecard modeling holds considerable importance due to its intrinsic quality as an explainable algorithm. Unlike certain complex algorithms that are often categorized as "black box" solutions, where the decision-making process is difficult to interpret or elucidate, logistical regression offers transparency and comprehensibility.

In the realm of credit scoring, where the assessment of risk factors is of utmost significance, being able to understand and interpret the variables and their impact on the final outcome is crucial. Logistic regression provides this advantage by generating coefficients that explicitly indicate the direction and magnitude of influence each variable holds. This transparency not only aids in understanding the underlying mechanisms driving the model's predictions but also facilitates the validation of the model's appropriateness and reliability.

Furthermore, logistics regression's transparent nature enhances collaboration and communication between data analysts, domain experts, and stakeholders. The insights gained from the model can be effectively communicated, discussed, and refined, enabling better-informed decisions and strategies.

In contrast, complex black box algorithms may deliver high predictive accuracy, but their inner workings remain concealed. This opacity can lead to challenges in building trust, addressing regulatory concerns, and understanding the model's limitations or potential biases.

2.5.2 Cut-off

The concept of "cut-off" holds paramount importance in binary classification algorithms, as it determines the optimal threshold for separating the "good" and "bad" classes. In other words, the "cut-off" is the decision point that allows observations to be classified based on their probability of falling into either of the categories.
The optimization of the "cut-off" directly impacts the performance of the classification model. If the "cut-off" is set too low, the model might be overly conservative, excessively categorizing observations as part of the majority class (e.g., "good" in the case of credit scoring models). Conversely, if the "cut-off" is set too high, the model could lack sensitivity, inadequately classifying observations as part of the minority class (e.g., "bad" in the context of credit scoring). Consequently, the appropriate choice of the "cut-off" achieves a balance between the precision and recall of the model, based on the specific objectives of the application.
Optimizing the "cut-off" is often accomplished using evaluation metrics such as the Receiver Operating Characteristic (ROC) curve or the Lift curve. These metrics enable the analysis of model performance at different "cut-off" thresholds and the selection of the one that maximizes the desired performance. In summary, the notion of "cut-off" constitutes a key element in the construction and adjustment of binary classification models, ensuring informed and effective decision-

making in domains like credit scoring.

2.6 Scorecard

Predictive models play a crucial role in yielding probabilities that can be subsequently interpreted as either falling within the "good" or "bad" class. In the realm of credit scoring, for instance, these probabilities are pivotal in assessing the creditworthiness of individuals. An interesting facet here is the "ScoreCard" methodology, which operates uniquely by generating probabilities from a provided score. This score-based approach bridges the gap between the numerical output and the actionable insights required for decision-making. Consequently, the "ScoreCard" not only transforms scores into meaningful probabilities but also facilitates a more accessible comprehension of the predictive outcomes, thereby enhancing its practicality and applicability in real-world scenarios.

The scaling of a "ScoreCard" is employed to transform the output of a predictive classifier into a score that represents a specific ratio between the "good" and "bad" outcomes. This process enhances the ease of comprehension and interpretation of a "ScoreCard," particularly for non-expert users.
Desirable properties of a "ScoreCard" encompass:

- The total score is positive;
- Points assigned to each attribute within the "ScoreCard" are positive;
- Reference scores exist with specific "odds" ratios between good and bad outcomes;
- Score differences hold a consistent meaning across the entire scale.

Scaling can be implemented using various approaches. Below, we outline the linear scaling method.

Formula

$$Score = offset + Factor \times \log(odds) \tag{2}$$

with, $\log(odds)$ refers to the logarithm of the score computed using a classification algorithm;
and $Factor$ and $Offset$ are scaling parameters to be determined.

The $Factor$ signifies the number of points (y) required to increase the odds by a certain specific quantity (m), as defined below:

$$Factor = \frac{y}{\log(m)}$$

Note

Scaling does not impact the predictive strength of the "ScoreCard."