# Industrial project description (banking credit scoring)

11 October 2024

The questions are answered from the perspective of an **Analyst**.

## 1 Planning the industrial research project

1. The primary goal of this project is to develop a credit scoring model that accurately predicts the default risk of potential borrowers in the banking sector. To achieve this, I am going to formulate a statistical model that enhances prediction accuracy over existing models, thereby optimizing the bank's lending decisions and minimizing financial losses due to defaults.

2. Thus, to tackle this problem, it is needed to improve the bank's ability to assess credit risk through a more precise credit scoring system. The results will be illustrated by implementing the new model on historical data and comparing its performance with existing models using evaluation metrics such as the Receiver Operating Characteristic and Area Under the Curve (ROC-AUC) and F1 score. Visual representations like graphs and tables will demonstrate improvements in predictive accuracy and the model's effectiveness in distinguishing between high-risk and low-risk borrowers.

3. The historical data comprises records of previous loan applicants, each characterized by a set of features and their corresponding credit outcomes. Each applicant is represented by a feature vector $\mathbf{X}_i = [x_{i1}, x_{i2}, \ldots, x_{in}]$, where $x_{ij}$ denotes attributes such as age, income, employment status and etc. The target variable $Y_i$ indicates whether the applicant defaulted ($Y_i = 1$) or not ($Y_i = 0$). This dataset can be structured as a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ and a target vector $\mathbf{Y} \in \{0, 1\}^m$, with $m$ applicants and $n$ features.

4. The quality of the credit scoring model will be assessed by optimizing an error function. Specifically, there will be minimization of the logistic loss function for classification:

$$L(\theta) = \frac{1}{m} \sum_{i=1}^{m} \ell \left( Y_i, f(\mathbf{X}_i; \theta) \right),$$

where $\ell$ is the loss function, and $f(\mathbf{X}_i; \theta)$ is the predictive model parameterized by $\theta$. The objective is to find the optimal parameters $\theta^*$ that minimize $L(\theta)$:

$$\theta^* = \arg\min_{\theta} L(\theta).$$

Additionally, it is needed to minimize the False Positive Rate and maximize the True Positive Rate to enhance the model's reliability.

5. The project's feasibility is supported by the availability of extensive historical data and computational resources. The error analysis plan includes data splitting into training, validation, and test sets, implementing $k$-fold cross-validation to ensure model generalization, and conducting residual analysis to identify patterns in errors. Also, statistical tests and sensitivity analyses will be performed to evaluate and refine the model's performance.

6. For successful project implementation, high-quality data is required that is accurate, complete, and consistent. It's crucial to include relevant features that significantly predict credit risk and to preprocess the data by handling missing values, outliers, and normalizing it. A sufficient sample size is necessary to train and validate the model effectively.

7. The following statistical and machine learning methods will be employed: logistic regression, decision trees, random forests, support vector machines, and gradient boosting machines like XGBoost. Hypotheses involve testing whether the new model provides a statistically significant improvement over the current one. By estimating the probability of default $P(Y_i = 1 \mid \mathbf{X}_i)$ using logistic regression:

$$P(Y_i = 1 \mid \mathbf{X}_i) = \frac{1}{1 + e^{-\theta^T \mathbf{X}_i}},$$

and utilizing ensemble methods, I aim to create an optimal probability model that enhances prediction accuracy.

## 2 Research or development?

This project embodies both development and elements of research. While credit scoring models are well-established, applying cutting-edge machine learning algorithms and customizing models to the bank's specific data represents a technological advancement. The novelty lies in integrating big data analytics, implementing ensemble methods to reduce variance and bias, and developing new variables through feature engineering that better capture borrower behavior.

The impact on the field of knowledge includes advancing credit risk modeling and setting new benchmarks for predictive accuracy in the banking industry. There is potential for academic contributions through publishing findings in financial and statistical journals. The usefulness of this project extends to financial institutions by enabling more informed lending decisions and reducing default rates, thereby contributing to economic stability. Customers will benefit from fairer credit assessments, leading to better loan terms for reliable borrowers.