Developing an Enhanced Credit Scoring Model Using Machine Learning Techniques: Achieving Explainability and Interpretability Using Dalex

By

OLAMIDE AKANNI

Submitted to

**The University of Roehampton**

In partial fulfilment of the requirements

for the degree of

**Master of Science**

**in**

**Data Science**

**Declaration**

I hereby certify that this report constitutes my own work, that where the language of others is used, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions, or writings of others.

I declare that this report describes the original work that has not been previously presented for the award of any other degree of any other institution.

Olamide Akanni

12/08/2024

Sign: A signature on a white surface

Description automatically generated

**Acknowledgments**

I want to thank The University of Roehampton for providing the resources and a conducive environment to learning, and the program conveners for always being helpful and supportive. I would also like to thank my supervisor, Tauseef Tahir whose guidance and expertise were instrumental in shaping this project.

**Abstract**

Credit scoring models are an important part of financial decision-making today. They influence the approval of loans and interest rates for millions of people. Traditional approaches to this problem often fail in terms of transparency and accuracy, mainly for underrepresented groups. The purpose of the project was to devise an enhanced model for credit scoring using machine learning techniques while ensuring its interpretability and explainability via the Dalex library.

The literature review examined the evolution of credit scoring methods, from simple heuristics to complex algorithms. It highlighted the tension between predictive accuracy and model interpretability, emphasizing the growing importance of explainable AI in finance. Key works by Lessmann et al. [2] and Biecek [15] provided benchmarks for model performance and frameworks for interpretability.

The implementation used the UCI Machine Learning Repository-based "Default of Credit Card Clients" dataset. After substantial preprocessing and feature engineering, a random forest model was developed and integrated into the Dalex (Descriptive mAchine Learning EXplanations) framework to make the results more interpretable. The model was deployed on a cloud platform with a user-friendly interface for real-world applications.

The findings proved that the Random Forest model could achieve 72% class accuracy and an AUC-ROC of 0.71, so it held reasonable predictive power. The Dalex framework provided valuable insights into feature importance, with the number of default months and recent payment status emerging as the most critical predictors. Partial dependence plots and individual conditional expectation plots produced nuanced understandings of feature impact on default probability.

While the performance of the model was at par with traditional approaches, its real strength lay in striking a balance between accuracy and interpretability. This project thus provides a contribution to the larger debate on fair and transparent AI in financial services and delivers a practical application in explainable AI for credit scoring.

Table of Contents

[Declaration ii](#_Toc172195179)

[Acknowledgements iii](#_Toc172195180)

[Abstract iv](#_Toc172195181)

[Table of Contents v](#_Toc172195182)

[List of Figures vii](#_Toc172195183)

[List of Tables viii](#_Toc172195184)

[**Chapter 1 Introduction** 1](#_heading=h.3dy6vkm)

[1.1 Problem Description, Context and Motivation 1](#_heading=h.1t3h5sf)

[1.2 Aims 1](#_heading=h.4d34og8)

[1.3 Objective 2](#_heading=h.1ksv4uv)

[1.4 Legal Considerations 2](#_heading=h.44sinio)

[1.5 Social Consideration 2](#_heading=h.2jxsxqh)

[1.6 Ethical Consideration 3](#_heading=h.z337ya)

[1.7 Professional Consideration 3](#_heading=h.z337ya)

[1.8 Background 3](#_heading=h.44sinio)

[1.9 Report Overview 4](#_heading=h.2jxsxqh)

[**Chapter 2 Literature Review** 5](#_heading=h.3j2qqm3)

[2.1 Traditional Credit Scoring Models 5](#_heading=h.1y810tw)

[2.2 Machine Learning in Credit Scoring 5](#_heading=h.4i7ojhp)

[2.3 Explainable AI (XAI) in Finance 6](#_heading=h.1y810tw)

[2.4 Dalex Framework 6](#_heading=h.4i7ojhp)

2.4.1 Comparison with other XAI frameworks 7

[2.5 Regulatory and Ethical Consideration 7](#_heading=h.4i7ojhp)

[2.6 Technology Review 8](#_heading=h.4i7ojhp)

[2.7 Practical Implementation of XAI in Credit Scoring 8](#_heading=h.1y810tw)

[2.8 Summary of Outcome 9](#_heading=h.4i7ojhp)

2.8.1 Summary of Literature Review Outcome 9

2.8.2 Summary of Technology Review Outcome 10

[**Chapter 3 Methodology** 11](#_heading=h.2xcytpi)

[3.1 Data Set 11](#_heading=h.1y810tw)

3.1.1 Data Selection and Description 11

3.1.2 Data Preprocessing 11

3.1.3 Data Partitioning 11

[3.2 Exploratory Data Analysis 11](#_heading=h.1y810tw)

3.2.1 Visualisation Techniques 11

3.2.2 Correlation Analysis 12

[3.3 Model Design (DESIGN) 12](#_heading=h.1y810tw)

3.3.1 Algorithm Selection 12

3.3.2 Feature Engineering 12

3.3.3 Hyperparameter Tuning 12

[3.4 Explainability Frame Work 12](#_heading=h.1y810tw)

3.4.1 Dalex Integration 12

3.4.2 Explanation Techniques 12

[3.5 Model Evaluation (TESTING AND EVALUATION) 13](#_heading=h.1y810tw)

3.5.1 Performance Metrics 13

3.5.2 Confusion Matrix Analysis 13

3.5.3 Interpretability Assessment 13

[3.6 Technologies and Tools (TECHNOLOGIES AND PROCESS) 1](#_heading=h.1y810tw)3

3.6.1 Programming Environment 13

3.6.2 Data Visualisation 13

3.6.3 Version Control and Collaboration (PROJECT MANAGEMENT) 13

[3.7 Ethical Consideration 13](#_heading=h.1y810tw)

3.7.1 Fairness and Bias Mitigation 13

2.7.2 Privacy and Security 14

[**Chapter 4 Implementation** 16](#_heading=h.1ci93xb)

[4.1 System Design and Architecture 16](#_heading=h.1t3h5sf)

4.1.1 Data Collection and Preparation 16

[4.2 Exploratory Data Analysis (EDA) 18](#_heading=h.1ksv4uv)

4.2.1 Descriptive Analysis 18

4.2.2 Visualization and Insights 18

[4.3 Feature Engineering 18](#_heading=h.1y810tw)

4.3.1 Creation of New Features 19

4.3.2 Feature Selection 19

[4.4 Model Develpoment 19](#_heading=h.1y810tw)

4.4.1 Model Selection 19

4.4.2 Training and Testing 20

4.4.3 Model Evaluation 20

4.4.4 Model Tuning 21

4.4.5 Addressing Class Imbalance 21

[4.5 Model Interpretability 22](#_heading=h.1y810tw)

4.5.1 Implementation of Dalex Framework 22

4.5.2 Explanation of Dalex Function Used 22

4.5.3 Hosting on Render 23

4.5.4 User Interface Development 23

[4.6 Challenges and Solutions 24](#_heading=h.1y810tw)

4.6.1 Dealing with Class Imbalance 24

4.6.2 Feature Selection and Multicollinearity 24

4.6.3 Model Interpretability vs. Performance Trade-off 25

[**Chapter 5 Evaluation and Result** 26](#_heading=h.3whwml4)

[5.1 Data Set Description 26](#_heading=h.1t3h5sf)

5.1.1 Data Source and Overview 26

5.1.2 Data Structure and Features 26

5.1.3 Data Quality and Preprocessing Considerations 27

[5.2 Exploratory Data Analysis (EDA) 28](#_heading=h.1t3h5sf)

5.2.1 LIMIT\_BAL (Amount of given credit in NT dollars) 29

5.2.2 AGE(Age in Years) 29

5.2.3 No\_of\_Default\_month (Number of months defaulted) 29

5.2.4 Average\_percentage\_offset (Ratio of average payment amount to bill amount 29

[5.3 Analysis of Categorical Variables 30](#_heading=h.1t3h5sf)

5.3.1 Gender Distribution (SEX) 30

5.3.2 Educational Background (EDUCATION) 30

5.3.3 Marital Status (MARRIAGE) 30

5.3.4 Consecutive Default Patterns (Consecutive\_Default) 30

5.3.5 Target Variable: Default Payment Next Month 31

[5.4 Feature Selection 32](#_heading=h.1t3h5sf)

5.4.1 Correlation Analysis 32

5.4.2 XGboost Feature Importance 33

[5.5 Model Performance 34](#_heading=h.1t3h5sf)

5.5.1 Overall Accuracy 34

5.5.2 Precision, Recall, and F1-Score 34

5.5.3 ROC-AUC Analysis 35

5.5.4 Confusion Matrix 35

5.5.5 Comparison with Baseline Models 35

[5.6 Model Interpretability (Using Dalex) 36](#_heading=h.1t3h5sf)

5.6.1 Global Feature Importance 36

5.6.2 Partial Dependence Plot (PDP) 37

5.6.3 Individual Conditional Expectation (ICE) Plot 38

5.6.4 Breakdown Plots for Individual Predictions 40

[5.7 User Evaluation 41](#_heading=h.1t3h5sf)

5.7.1 The Deployment Model Access 42

[5.8 Related Work 4](#_heading=h.1t3h5sf)2

[**Chapter 6 Conclusion** 43](#_heading=h.3whwml4)

[6.1 Future Work 43](#_heading=h.2bn6wsx)

[6.2 Reflection 44](#_heading=h.qsh70q)

[References 46](#_heading=h.3as4poj) - 49

[Appendices 50](#_heading=h.1pxezwc)

[Appendix A: Project Proposal 51](#_heading=h.49x2ik5)

[Appendix B: Project Management 52](#_heading=h.2p2csry)

[Appendix C: Artefact/Dataset 53](#_heading=h.147n2zr)

[Appendix D: Screencast 54](#_heading=h.3o7alnk)

**List of Figures**

[1 Descriptive Statistical Dashboard for Selected Features 43](#_heading=h.2bn6wsx)

[2 Dashboard of Categorical Variables Distributions 44](#_heading=h.qsh70q)

[3 Heatmap Showing correlation Analysis 43](#_heading=h.2bn6wsx)

[4 XGboost Feature Importance 44](#_heading=h.qsh70q)

[5 Evaluation Metrics of the Model 43](#_heading=h.2bn6wsx)

[6 Global Feature Important (variable ..Y axis.. ..loss on the X axis) 44](#_heading=h.qsh70q)

[7 PDP for NO\_of \_Default\_Month 43](#_heading=h.2bn6wsx)

[8 PDP for PAY 44](#_heading=h.qsh70q)

[9 ICE Plot for No\_of Default\_Month 43](#_heading=h.2bn6wsx)

[10 ICE Plot for PAYO 44](#_heading=h.qsh70q)

[11 Break Down – High Risk Individual 43](#_heading=h.2bn6wsx)

[12 Beak Down – Low Risk Individual 44](#_heading=h.qsh70q)

[13 Credit Defualt Dep 44](#_heading=h.qsh70q)

**List of Tables**

[1 Literature Review Outcomes 43](#_heading=h.2bn6wsx)

[2 Technology Review Outcomes 44](#_heading=h.qsh70q)

[3 Data Structure and Feature Description 43](#_heading=h.2bn6wsx)

# **1. Introduction**

One of the most important applications in the modern theory of finance is credit scoring models, which define whether to accept or decline lending applications and interest rates for millions of people around the world. Classic approaches, like logistic regression and decision trees, are standard tools to solve this kind of problem but often are black box and not accurate for people with thin files or for underrepresented classes [1]. This paper is focused on the construction of an improved credit-scoring model using machine-learning techniques that ensure interpretability and explainability with the Dalex library.

## 1.1 Problem Description, Context, and Motivation

Credit scoring models are crucial for access to financial products, yet traditional and innovative models often remain non-transparent and non-interpretable, thus challenging the industry, regulators, and consumers [2]. This opacity is common, especially when regulatory scrutiny increases and consumers demand fairer decisions. Traditional models often miss the subtlety of modern financial behavior, while more advanced machine-learning models act as a "black box" [3].

The aim of this research is to construct accurate, yet interpretable credit scoring models that will transform the industry toward fostering equal credit access, enhancing risk management, and ensuring transparency within the financial system [5].

## 1.2 Aims

1. Develop an advanced credit scoring model using cutting-edge machine learning techniques for higher accuracy than traditional models.
2. Utilize the Dalex framework to ensure the model's explainability and interpretability, addressing the "black box" issue common in complex machine learning models.
3. Contribute to the broader conversation on fair and transparent AI in financial services by showcasing a practical application of explainable AI in credit scoring.

## 1.3 Objectives

1. Conduct an in-depth review of current credit scoring methodologies and machine learning techniques relevant to credit risk assessment.
2. Gather and preprocess relevant datasets to prepare for model training.
3. Develop, train, and validate a machine learning model for credit scoring, selecting the best-performing one.
4. Utilize the Dalex framework for explainability, ensuring model transparency and analyzing feature importance and model behavior.

## 1.4 Legal Considerations

The legal frameworks within which the credit scoring models should move are very complex, with a myriad of consumers' rights and policies that ensure equity in lending. In the U.S., it is regulated by Fair Credit Reporting Act (FCRA) and Equal Credit Opportunity Act (ECOA), which oversees and regulates credit scoring to ensure no discrimination or unfair practices take place [6]. Internationally, the high bar for data protection and privacy regarding automated decision-making processes was set by the EU's General Data Protection Regulation (GDPR) [7].

This is where model interpretability through a Dalex integration comes in and ensures compliance with transparency laws. Precise explanations for model decisions prove conformance to both the intent and specifics of these very laws [8]. During development and deployment, strict adherence to these legal frameworks reduces regulatory risks and furthers the objective to establish a fair, transparent, and reliable credit scoring system.

## 1.5 Social Considerations

Models for credit scoring have huge implications for access to financial resources and, therefore, far-reaching impacts on society. For that matter, models should be designed not to perpetuate the biases against certain demographic groups. We want to have a model that is transparent and hence trustworthy, which would result in an equable financial system. This work improves accuracy and interpretability in credit risk assessment while maintaining robust risk assessment, thereby promoting financial inclusion and ultimately contributing towards a more inclusive and varied financial system [9].

## 1.6 Ethical Considerations

The credit scoring model therefore has deep ethical dimensions: fairness, accountability, and transparency—some of the core elements of the responsible development of AI in finance [10]. Considering AI can easily replicate biases, the model mitigates these discriminatory outcomes across demographic groups. Rigorous protocols was implemented on the detection and management of bias in the training data and model decisions, especially with regular audits on datasets. The framework, Dalex, for explainable AI thus supports our imperative of transparency, including giving clear explanations of credit decisions [11]. The ethical clearance was obtained via a holistic review process that showcases an effort in ensuring data privacy, mitigation of bias, and interpretability of the model. A look at an efficient credit-scoring system with top ethics.

## 1.7 Professional Considerations

The project adheres to Responsible AI Development principles recommended by the IEEE and ACM, working on transparent and interpretable machine learning models targeting financial decision-making. Subscription to the best practices for data science and financial modelling in terms of integrity, reproducibility, and transparency during the research process, entailing meticulous documentation throughout and rigorous testing in order for model capabilities and limits to be evidently expressed.

With this, the research would like to contribute very substantively to the professional discourse about responsible AI in finance—the one aiming at the balance of the most sophisticated predictive capabilities of our credit scoring system with its full explainability. The work has the potential to influence industry standards and foster more ethical and transparent AI systems in financial services. More fundamentally, we believe this project offers us the chance to demonstrate AI's potential for establishing fairer and more inclusive financial systems—establishing new standards regarding responsible innovation in credit scoring.

## 1.8 Background

Starting from very simple heuristics to sophisticated statistical models, credit scoring has moved in step with increasing demand from the financial sector for accurate risk assessment within increasingly complex data-rich environments [12]. Traditional methods—logistic regression and decision trees—are interpretable but poorly capture the complexity of credit risk.

Advanced machine learning techniques—such as random forests, gradient boosting machines, and neural networks—have opened new frontiers in credit scoring accuracy by detecting complex nonlinear relationships that exist in the data [13]. However, it is exactly these "black box" models which often show no transparency, leading to tension between accuracy and explainability [14].

One field of research that has recently grown out of this challenge is explainable AI (XAI). Tools like Dalex [15] offer frameworks to create transparent explanations for the often complex decisions reached by a model. This work uses the potential offered by Dalex in building a credit scoring model that balances predictive accuracy with interpretability, and therefore serves the interests of the industry while being aligned with the trends toward responsible AI in critical decision-making processes.

## 1.9 Report Overview

The report explores the development of an explainable credit scoring model using machine learning and the Dalex framework, progressing logically from problem identification to evaluation.

Chapter 1: The Introduction, Outlines the research context, defines the problem, and details aims and objectives, addressing legal, social, ethical, and professional considerations.

Chapter 2: Literature Review examines credit scoring methodologies, traditional approaches, machine learning integration in finance, and recent developments in explainable AI, providing foundational knowledge.

Chapter 3: Methodology explains the chosen approach, including machine learning algorithm selection and Dalex implementation for model interpretability, aligning with research objectives and transparency goals.

Chapter 4: Implementation details data preprocessing, model training, and application of explainability techniques, discussing practical challenges and solutions.

Chapter 5: Results evaluates model performance in terms of predictive accuracy and interpretability, with analysis of model outputs, Dalex insights, and comparative analysis against existing approaches.

Chapter 6: Conclusion synthesizes key findings, reflects on objectives, discusses credit scoring implications, and suggests future research directions.

# 

# **2. LITERATURE REVIEW**

Credit scoring is vital for financial decision-making, impacting loan approvals, interest rates, and risk management. The evolution from simple heuristics to complex algorithms mirrors the growing complexity of financial environments and consumer behaviour. This review examines the current credit scoring methods, emphasizing the challenge of balancing predictive accuracy with model interpretability and transparency [16].

## 2.1 Traditional Credit Scoring Models

Traditional credit scoring relies on interpretable models that offer clear insights into decision-making processes. Logistic regression, a cornerstone in credit risk assessment, provides a straightforward relationship between input variables and outcomes [12]. However, its dependence on linear relationships limits its ability to capture complex patterns in modern financial data [17].

Decision trees present credit decisions in an intuitive flowchart structure but are susceptible to overfitting and instability with complex datasets [18]. Credit scorecards, derived from logistic regression models, offer a simple scoring system but may struggle to adapt to dynamic financial behaviors [19].

While these traditional models are transparent, they often fail to capture the intricacies of today's financial landscape, potentially leading to suboptimal predictive performance.

## 2.2 Machine Learning in Credit Scoring

Machine learning has, to a great extent, improved the predictive accuracy in credit scoring. Ensembling methods, such as random forests, improve the prediction accuracy at the cost of increased complexity by averaging several decision trees [20]. Gradient boosting machines build sequential models to improve their performance but often sacrifice interpretability [21]. The use of neural networks, more so deep learning models, in modeling complex relationships that exist in large data makes them very effective in credit risk assessment. However, usually, this increased accuracy is attained at the cost of interpretability, generally rendering them "black boxes" [22]. Random forests have become very powerful in credit scoring because they can handle complex and nonlinear relationships fairly well and show robustness to overfitting [20]. Brown and Mues [24] showed that Random Forests outperformed all other methods in most cases, specifically on an imbalanced dataset. Lessmann et al. [2] presented the detailed benchmark study of 41 classifiers using eight credit scoring datasets, where it was proved that machine learning methods significantly outperform the traditional ones based on statistical methodologies. Xia et al. [25] conducted a comparison between XGBoost, random forests, and deep neural networks and reported the best overall performance of the XGBoost algorithm. The tension between model performance and interpretability remains high in a field like finance, where the justifications for decisions are paramount for regulative compliance and stakeholder trust. Byanjankar et al. [26] commented that methods should be chosen considering predictive performance, interpretability, and the requirements on computational resources when making decisions, specifically in regulated industries.

## 2.3 Explainable AI (XAI) in Finance

Demands for interpretability in credit scoring have been on the rise, which has precipitated the development related to explainable AI. Techniques like LIME and SHAP are model-agnostic methods that provide methods for post hoc explanations of complex models.

LIME works by approximating the black-box model at the local level by using an interpretable model for explaining individual predictions, ensuring better transparency but at high computational cost [27]. SHAP values provide a unified measure of feature importance based on cooperative game theory that offers consistency and interpretability, though it is computationally intensive [28].

These methods of XAI primarily have the goal of closing the gap between high performance by state-of-the-art machine learning models and the requirements for transparent, justifiable decisions. However, even in those cases, challenges still remain as to how to make those technical explanations meaningful for non-technical stakeholders [29].

## 2.4 Dalex Framework

The Dalex framework, introduced by Biecek [15], has gained significant traction in explainable AI, particularly in finance. Dalex provides a unified interface for various model-agnostic explanation methods, facilitating implementation and comparison of different interpretability techniques.

Dalex offers comprehensive tools for explaining and diagnosing machine learning models [15], integrating various XAI techniques into a cohesive framework supporting multiple machine learning libraries. Compared to standalone tools like LIME and SHAP, Dalex provides a more integrated approach to model explanation. Its application in finance has been notable, where understanding model behaviour is crucial for compliance and trust [30].

Key capabilities of Dalex include:

1. Variable importance plots
2. Partial dependence plots
3. Individual conditional expectation plots
4. Break-down plots for individual predictions
5. Model performance and residual diagnostics

Despite some computational challenges, Dalex's comprehensive nature makes it well-suited for credit scoring applications where explanations must be understandable to a wide range of stakeholders.

### 2.4.1 Comparison with Other XAI Frameworks

While Dalex provides an overall suite of tools, other XAI frameworks like SHAP [28] and LIME [27] are equally common for the task. Combining cooperative game theory, SHAP creates locally consistent and accurate values of feature importance. Contrasted with LIME, that will look for local interpretability through local approximations of the model at hand. Dalex offers a far more integrated approach, as it packs several explanation methods into one framework, greatly easing comparisons between techniques. However, the choice of an XAI framework mostly depends on the problem requirements and the model being explained [31].

## 2.5 Regulatory and Ethical Considerations

Models for credit scoring are in regulators' focus worldwide. The European Union's GDPR establishes a "right to explanation" of automated decisions [8], while a variety of US regulations require that credit reporting be fair and transparent [6]. Ethics-based fairness or non-discrimination and absence of bias in AI-driven credit scoring are crucial tools in preventing discrimination and ensuring equal access to credit [32], [33].

Some of the proposed techniques for handling fairness challenges range from preprocessing techniques [34] by, to adversarial debiasing methods [35]. There are a number of standards and best practices from the financial industry that include model interpretability and human oversight, with clear guidelines from the European Banking Authority [36] and the Federal Reserve [37].

The debate on inherently interpretable models versus post-hoc explanations for complex models also persists. Rudin [38] makes a case for inherently interpretable models in high-stakes decisions, though others suggest careful design of explanation techniques suffices for transparency [27].

These best practices include frequent model validation, comprehensive documentation, fair representation within the training dataset, fairness metrics, and human oversight. The choice of approach depends on the requirements of a method and regulatory constraints in use, along with data complexity.

## 2.6 Technology Review

Explainable credit scoring models rely on diverse technologies. Machine learning libraries like Scikit-learn[[1]](#footnote-1) and TensorFlow[[2]](#footnote-2) provide robust tools for building and training models [39], [40]. Scikit-learn, built on NumPy,[[3]](#footnote-3) SciPy[[4]](#footnote-4), and matplotlib[[5]](#footnote-5), is widely adopted in credit scoring applications due to its user-friendly interface and comprehensive tools [39].

Scikit-learn's implementations of algorithms like Random Forests[[6]](#footnote-6), Gradient Boosting[[7]](#footnote-7), and Logistic Regression[[8]](#footnote-8) are frequently used in credit scoring. Its consistent API design allows easy comparison of different models. The library's preprocessing tools and model selection modules are crucial for preparing financial data and tuning models.

Effective data preprocessing is vital for model performance [41]. For deployment, cloud platforms like AWS[[9]](#footnote-9), Azure[[10]](#footnote-10), and Google Cloud[[11]](#footnote-11) ensure scalability and reliability [42].

This project will implement the credit scoring model using Python, with Scikit-learn as the primary machine learning framework and Dalex for model explainability, balancing performance, interpretability, and operational feasibility.

## 2.7 Practical Implementation of Explainable AI in Credit Scoring

Case studies demonstrate successful implementation of explainable AI in credit scoring. Bracke et al. [16] used SHAP values to explain predictions of a machine learning model for mortgage risk, providing individual-level explanations. Sirignano et al. [43] developed a deep learning model for mortgage risk with interpretability measures using partial dependence plots and variable importance measures.

Challenges in implementing explainable models include:

1. Balancing model complexity with interpretability
2. Ensuring consistency of explanations
3. Translating technical explanations for non-technical stakeholders
4. Meeting regulatory requirements for model transparency

Best practices for model transparency in financial services include:

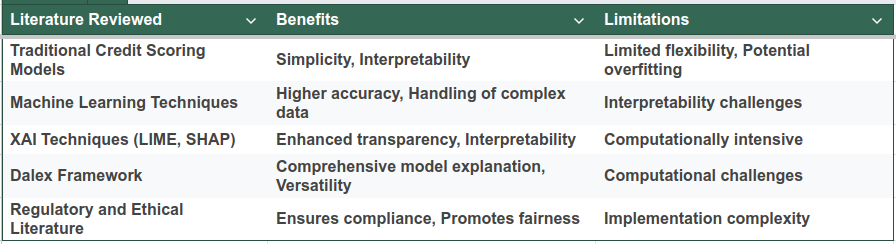
1. Providing clear, consistent explanations for model decisions
2. Ensuring actionable explanations for stakeholders and customers
3. Regularly auditing models for fairness and bias
4. Maintaining human oversight and override capability
5. Documenting the model development process

## 2.8 Summary of Outcomes

The literature review highlights the tension between the predictive power of advanced machine learning models and the need for interpretability in credit scoring. Traditional models offer transparency but may fall short in capturing the complexities of modern financial data. In contrast, sophisticated ML algorithms demonstrate superior predictive performance but often lack interpretability.

### 2.8.1 Summary of Literature Review Outcomes

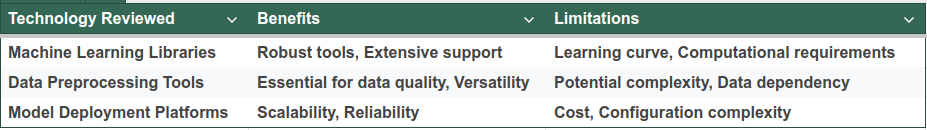
#### Table 1: Literature review outcomes



The literature review outcomes table critical analyse and shows that while advanced machine learning techniques offer improved performance, they require robust explainability tools like Dalex to ensure transparency and compliance.

### 2.8.2 Summary of Technology Review Outcomes

#### Table 2: Technology review outcomes



The technology review outcome table underscores the necessity of selecting appropriate tools and platforms to balance performance, interpretability, and operational feasibility.

# 

# **3. Methodology**

## 3.1 Dataset

### 3.1.1 Dataset Selection and Description

The "Default of Credit Card Clients" dataset from the UCI Machine Learning Repository was selected, containing 30,000 records with 24 features each. The target variable indicates credit card payment default in the subsequent month, with features covering demographics, credit history, and payment behaviors.

### 3.1.2 Data Preprocessing

Missingness Analysis

The dataset was examined, and missing values were imputed using mean/median/mode, k-nearest neighbours and regression imputations techniques to ensure data integrity and avoid biases.

Descriptive Analysis:

Descriptive statistics, such as means and standard deviations for continuous variables, were computed to understand their central tendencies and variability. The distributions of categorical variables were assessed in search of patterns and outliers that would likely affect model performance.

Handling Categorical Variables:

To incorporate categorical variables into the model, dummy encoding was used. This process converts categorical data into binary (0/1) variables, making them suitable for analysis in machine learning algorithms without introducing biases from arbitrary numerical encoding.

### 3.1.3 Data Partitioning

The dataset was manually divided into training and test sets in an 80:20 ratio to avoid bias in target class representation. This makes the evaluation metric very simple yet effective. Its 80-20 split was decided upon because this provides a good trade-off between training the model on enough data and saving enough for its precise evaluation. In this case, N-fold cross-validation is avoided due to the excess computational overhead and that the dataset is big and well-distributed enough that the risks of overfitting are minimal, and thus a single split will suffice for a reliable evaluation. This approach was chosen to maintain control over the balance of default and non-default instances, unlike automatic methods that might not account for this balance.

## 3.2 Exploratory Data Analysis

### 3.2.1 Visualization Techniques

Several visualization techniques provided insight into the data. Histograms visualized the distribution of continuous variables, detecting skewness and outliers. Pareto plots highlighted the frequency distribution of categorical variables, while the frequency distribution table summarized occurrences in binary data.

### 3.2.2 Correlation Analysis

A Spearman's rho correlation heatmap was used to visualize the relationships between variables. Since it is robust against skewed data, a heatmap has been chosen that governed important inter-variable relationships and pointed out impactful features

## 3.3 Model Design (DESIGN)

### 3.3.1 Algorithm Selection

Machine learning algorithms, including Random Forest, XGBoost, and Neural Networks, were considered. Random Forest was selected for its balance of performance and interpretability, robustness to overfitting, and ability to handle many features [20,27]

### 3.3.2 Feature Engineering

New features have been engineered based on domain knowledge. Thereafter, techniques in recursive feature elimination and mutual information selected the most relevant features, hence reducing dimensionality and improving performance as well.

### 3.3.3 Hyperparameter Tuning

It was performed using grid search to optimize the Random Forest settings of hyperparameters. Cross-validation ensured model generalizability and avoided overfitting..

## 3.4 Explainability Framework

### 3.4.1 Dalex Integration

The Dalex framework was integrated with the Random Forest model for enhanced interpretability, providing comprehensive tools to explain model predictions [15].

### 3.4.2 Explanation Techniques

Variable importance plots showed influential features, partial dependence plots showed the relationship between the features and outcomes, and individual prediction explanations offered an in-depth case insight.

## 3.5 Model Evaluation (TESTING AND EVALUATION)

### 3.5.1 Performance Metrics

The model performance was tested using the following criteria: accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve, which measures the model's ability to differentiate between default and non-default cases.

### 3.5.2 Confusion Matrix Analysis

A confusion matrix analyzed classification performance in relation to true positives, true negatives, false positives, and false negatives, giving details of the strengths and weaknesses of the model.

### 3.5.3 Interpretability Assessment

The quality of Dalex explanations was assessed and compared with traditional credit scoring methods, evaluating the model's practical applicability and transparency.

## 3.6 Technologies and Tools (TECHNOLOGIES AND PROCESSES)

### 3.6.1 Programming Environment

Python was used, drawing on its machine learning and data analysis libraries. Notable ones were: scikit-learn - model development, pandas - data manipulation, and Dalex for interpretability. [39].

### 3.6.2 Data Visualization

Matplotlib and Seaborn provided robust tools for creating informative and visually appealing plots, crucial for understanding data distributions and model outputs.

### 3.6.3 Version Control and Collaboration (PROJECT MANAGEMENT)

In this project, a Gantt chart was utilized to schedule and manage the timeline of various project phases effectively. Project reporting and documentation were streamlined using Trello, facilitating the organisation, supervising, and tracking tasks. For version control, Git was employed, ensuring that code changes were systematically managed. The project's codebase was hosted on GitHub, providing a centralized repository for collaboration and code storage.

## 3.7 Ethical Considerations

### 3.7.1 Fairness and Bias Mitigation

Techniques to detect and mitigate biases, like reweighting or adversarial debiasing, ensured model fairness across demographic groups [32].

### 3.7.2 Privacy and Security

Data anonymization techniques protected individuals' privacy. Compliance with GDPR and FCRA ensured data privacy and security [6]

# **4. Implementation**

The following section provides the practical implementation of the credit scoring model based on the methodologies outlined in the previous chapter.

### 4.1 System Design and Architecture

The procedures have a designed a modular, scalable Credit Scoring System which is composed of a multitude of interdependent components:

* Module for Data preprocessing: Cleansing, transforming, and providing raw data in forms that are convenient to analyze and model.
* Feature engineering module: This module focuses on the generation of new features and the selection of relevant attributes.
* Model Training Module: According to Random Forest algorithm integration, which will be responsible for handling the training and developing process.
* Evaluation Module: This module provides estimates for the model's performance or accuracy and outputs the results based on its metric measurements.
* Interpretability module: Employs the use of the Dalex framework to interpret the predicted values.
* User Interface: A web interface through which client data can be keyed in in order to obtain predicted credit risk.

The system was predominantly implemented in Python, using its various libraries to data science and machine learning problems. Some of the important technologies and tools are mentioned below:

* Python 3.10: Main programming language for the whole Implement EOF.
* Pandas 2.1.4: for data manipulations and analyses.
* Scikit-learn 1.3.2: This is for the implementation of the model and several preprocessing techniques.
* Matplotlib 3.7.1. and Seaborn0.13.1: For data visualization.
* XGBoost 2.1.0: Applied during feature selection.
* Dalex 1.4.1: For model interpretation and explanation.
* Flask 2.2.5: For creating the web application backend.
* React 17.0.2: User interface development.
* Render: To host and deploy the web application.

This architecture ensures concern separation clearly, whereby making the system maintainable and easily extendable in the future.

## 4.1.1 Data Collection and Preparation

Data Acquisition:

This data taken from the UCI Machine Learning Repository is the "Default of Credit Card Clients" dataset. This dataset is from the records of a Taiwanese bank's credit card clients. The defined number of instances in the dataset is 30,000, where every instance represents an individual credit card client. Each instance is described by 24 attributes, representing information containing demographic details, credit data, payment history, and bill statements.

The main feature of this dataset is its relevance: issues of most importance in the themes of research into credit scoring are represented within it, and, at the same time, it deals with a great number of potentially influential credit management factors all together. Thereby, it provides a rich base for the development and testing framework of a credit scoring model.

Data Preprocessing:

The important stage in getting the data set ready for the analysis and modeling process.

*Data Cleaning:*

* An initial inspection was done to check for missing data
* Duplicate entries were checked for and removed, ensuring data integrity.

*Feature Normalization and Standardization:*

For numeric features such as LIMIT\_BAL, AGE, and BILL\_AMT, scaling was performed using StandardScaler from scikit-learn. This ensured that all features were on a comparable scale, so no dominance would ensue during model fitting between features which have large magnitudes.

*Addressing Class Imbalance:*

* The target variable (default payment next month) exhibited significant class imbalance, with approximately 78% non-defaulters and 22% defaulters.
* An under sampling technique was used to reduce this imbalance, which randomly removed two-thirds of the instances of non-defaulters, thus bringing balance to the imbalanced data with a ratio of about 40:60 of defaulters and non-defaulters for further usage.
* This approach was chosen over oversampling techniques because of the added synthetic data points that may lead to overfitting.

*Multicollinearity Assessment:*

* A correlation analysis was performed to identify highly correlated features.
* Variables with a correlation index equal to or exceeding 0.8 weredroped to solve multicollinearity.
* Features were therefore deleted largely based on domain knowledge and their potential importance in the credit risk assessment.

*Feature Engineering:*

New features were created to capture additional insights:

1. Consecutive\_Default: A binary feature indicating if a client defaulted in consecutive months. 1 for Yes and 0 for No
2. No\_of\_Default\_month: Overall number of months the client defaulted, and the values are numeric.
3. Average\_percentage\_offset: Average amount paid over a six-month period relative to average average bill amount. Values are numerical.

## 4.2 Exploratory Data Analysis (EDA)

The characteristics and underlying patterns of the dataset were detailed during the explorative data analysis phase. So many statistical techniques and visualization methods have been adopted, focusing on variables both continuous and categorical, to extract information that may be influential for the downstream building of models.

### 4.2.1 Descriptive Analysis

We described in detail continuous variables such as age, credit limit, and bill amounts. It involved the estimation of central tendency measures: mean, median, and mode, and dispersion measures: standard deviation, range, and interquartile range. Skewness and kurtosis have also been computed, which enabled us to remark on the shape of the different distributions. These statistics gave us useful insights into the nature of our data and probable outliers. For instance, the distribution for credit limit is positively skewed so that the bulk of the data is condensed toward lower credit limits with a tail stretching to higher values.

It computed the frequency distribution for these categorical variables, including gender, education level, and marital status, along with proportions. This gave insight into some of the key demographic attributes of the credit-card client base.

### 4.2.2 Visualization and Insights

Visualization played a key role in our EDA process. A series of histograms for continuous variables to visually assess their distributions and identify any unusual patterns or outliers. Bar charts were employed for categorical variables to illustrate the relative frequencies of different categories. These visualizations were instrumental in identifying trends and patterns that might not be immediately apparent from numerical summaries alone.

A correlation heatmap was generated to visualize the relationships between numerical variables. This heatmap was particularly useful in identifying potential multicollinearity issues and highlighting features that showed strong correlations with the target variable (default payment next month).

## 4.3 Feature Engineering

The feature engineering process was designed to extract more valuable information from the existing dataset by creating new features that increase the predictive power of a model. In this phase, new features were created, and at the same time, feature selection was done with respect to the most relevant ones in developing a model.

### 4.3.1 Creation of New Features

During the EDA process, a numerical feature was calculated based on the sum of the number of times a client had defaulted in each of the last months that were reported to give a historical perspective on the client's financial behavior over the last six months. A ratio feature was computed that represented the average payback amount relative to the bill amount over a period of six months, giving a rough heuristic assessment of the payback behavior of a client in relation to credit use.

### 4.3.2 Feature Selection

We further refined the set of features using another feature importance capability of the XGBoost algorithm. An instance of the XGBoost model was trained on the dataset, and then it was evaluated for the relative importance of each feature with respect to the predictions made by that model. Variables having a feature importance greater than 1 were selected

## 4.4 Model Development

The critical steps in this model development phase include model selection, data splitting, model training, evaluation, and tuning.

### 4.4.1. Model Selection

After considering several algorithms, Random Forest was chosen as the key modeling technique for this project. This was based on the following factors: first, ensemble nature—this technique has combled many decision trees into a single model for high predictive accuracy; second, it provides good balancing in both model complexity and interpretability. This became very important since our prime focus is model explainability. Moreover, Random Forest's capacity for nonlinear relations and interactions of features put it well in line with the challenge of credit risk prediction.

### 4.4.2 Training and Testing

The dataset was then split into two portions: 80% for training and 20% for testing. This was done so that there would be enough data to significantly train the model but also have enough data left over to test for any biases and to have reliable evaluations on the performance on unseen data. It was stratified on the target variable to ensure that the split had basically the same proportion of defaulters and non-defaulters as in the whole dataset.

Model training was done with the scikit-learn library in Python. An instance of the RandomForestClassifier class with initial hyperparameters set by common best practices and knowledge about the domain was created. Then, using the fit() method, it was fitted on the training data. In the process, Random Forest algorithm builds an ensemble of decision trees which are trained on a bootstrap sample of the training data and at every split consider random subset of features.

### 4.4.3 Model Evaluation

Model performance was evaluated regarding accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic curve. All these metrics vary in what they point to about the model performance. Accuracy provided a general sense of correct predictions, while precision and recall gave insights into how the model is performing with regard to positive cases of default. The F1-score provided a balanced measure between precision and recall, with AUC-ROC testing how well the model can distinguish between classes at many threshold settings.

A confusion matrix was also designed and analyzed to provide a breakdown of the model's predictions, thus helping in understanding how the true positives, true negatives, false positives, and false negatives are arranged. This analysis was particularly important given the imbalanced nature of our dataset and the differing costs associated with different types of misclassifications in credit scoring.

### 4.4.4 Model Tuning

Model tuning was an iterative process geared toward raising the performance of the model. This included changing such hyperparameters as the number of trees that will make up the forest, the maximum depth a tree is allowed to grow, and the minimum number of samples required to possibly split an internal node. We followed grid search combined with cross-validation to go through space in a systematic way for all possible hyperparameters in order to find the best configuration.

### 4.4.5 Addressing Class Imbalance

In order to handle large class imbalance in our dataset, an undersampling technique for the majority class has been implemented. More precisely, we removed randomly two-thirds of non-defaulter instances from the training set. This has been preferred over approaches of oversampling to avoid synthetic data points that would cause overfitting. The undersampling was done only on the train data to keep the test set intact for unbiased evaluation. This helps in getting a model evaluation that will be representative of the real class distribution of the population.

The choice of this step—undersampling over oversampling in this particular case—had been guided by several reasons:

* Data Integrity: Undersampling will not introduce synthetic data points and retains the characteristic of the minority class.
* Computational Efficiency: By reducing the dataset size, it was faster to train and iterate through models.
* Avoiding Overfitting: Sometimes, especially in the case of synthetic examples generated by oversampling, this can result in overfitting.

The last ratio, after undersampling, was approximately 40:60 for defaulters to non-defaulters, which would make the representation of both classes with regard to model training almost equitable.

## 4.5 Model Interpretability

One of the basic objectives of this study has been to increase model interpretability, given the relevant transparency of credit-scoring models. We will ensure the use of the Dalex framework to provide an overall explanatory tool on a machine-learning model.

### 4.5.1. Implementation of Dalex Framework

Integration of the Dalex framework was done after the Random Forest model had been trained and validated. The implementation steps were:

* Explainer Object Creation: We created a Dalex explainer object, including our fitted random forest model, the feature set, and the target variable.
* Data Preparation: We prepared a subset of the test data with the objective of generating explanations to ensure that we had a diversity of instances for interpretation.
* Function Integration: We integrated a host of Dalex functions, which extended at once our analysis pipeline with the functionality needed to perceive model behavior and decision making from many diversified perspectives.

### 4.5.2 Explanation of Dalex Functions Used

Several key features of functions within Dalex were harnessed in an attempt to gain holistic model interpretability:

* model\_parts(): The function was employed to calculate global feature importance: It was through this that we got to know which feature contributed more to the response variable globally, across the entire data set.
* model\_profile(): This function helped us create Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) plots. These plots depicted average and individual response changes upon changes in feature.
* predict\_parts(): This function enabled the creating of Break Down plots for single predictions. By using it, we could get an augmented version of one example explaining the reasons for the model's decision for a given credit applicant.

The Dalex functions provided a versatile approach to model interpretation, at both the global model level and at the level of individual predictions.

The final phase of our project involved deploying the trained and interpreted model into a production environment, making it accessible for real-world credit scoring applications.

### 4.5.3 Hosting on Render

The model deployment was done Render, which is a cloud platform particularly known for perfect deployment and scaling. Deployment required us to do the following:

* Model Serialization: The trained Random Forest model was serialized using Python's pickle library, which will enable it to load with much ease in the deployment environment.
* API Development: We have developed a Flask-based API that will process the incoming requests, sending these through our model as processed inputs, returning the predictions.
* Setting up the Environment: All the required environment variables and dependencies were set in the Render/Vercel platform for the smooth running of our application.
* Continuous Integration/Continuous Deployment: We set up a CI/CD pipeline that automated the deployment process, ensuring any updates to the model or API can quickly and reliably be pushed to production.

### 4.5.3 User Interface Development

user interface that allows end users to easily use the developed scoring model to score credits. We implemented the interface using the React library of JavaScript, a popular library in developing interfaces. The main features implemented in the interface include:

* Input Form: A sheet on which the users could input relevant information of applicants for credit, as available in our model.
* Real-time prediction: This is when the form can be filled and submitted and a prediction on credit default risk is given promptly.
* Explanation Display: Integrate Dalex explanations to give insights to the users into the prediction-driving factors.
* Responsive design: ensuring the interface is accessible and works on various devices and screen sizes.

## 4.6 Challenges and Solutions

Throughout the implementation process, we encountered several challenges that required innovative solutions:

### 4.6.1 Dealing with Class Imbalance

As mentioned earlier, the significant class imbalance in our dataset was a major challenge. Our solution of undersampling the majority class proved effective, but it required careful consideration to ensure we didn't lose important information from the non-defaulter class. We mitigated this risk by:

* Performing multiple random undersamples and averaging model performance to ensure robustness.
* Carefully monitoring performance metrics, particularly recall for the minority class, to ensure the model remained effective at identifying potential defaulters.

### 4.6.2 Feature Selection and Multicollinearity

Another significant challenge was selecting the most relevant features while addressing multicollinearity issues. We approached this by:

* Combining correlation analysis with XGBoost feature importance to get a comprehensive view of feature relevance and redundancy.
* Iteratively removing highly correlated features and re-evaluating model performance to find the optimal feature set.

### 4.6.3 Model Interpretability vs. Performance Trade-off

Balancing the need for model interpretability with achieving high predictive performance was an ongoing challenge. We addressed this by:

* Opting for Random Forest, which offers a good balance between performance and interpretability.
* Leveraging the Dalex framework to enhance the interpretability of our complex model without sacrificing predictive power.

This comprehensive implementation lays a strong foundation for the subsequent results and analysis, where we will delve into the specific outcomes, performance metrics, and insights gained from our credit scoring model.

# **5. Evaluation and Results**

## 5.1 Dataset Description

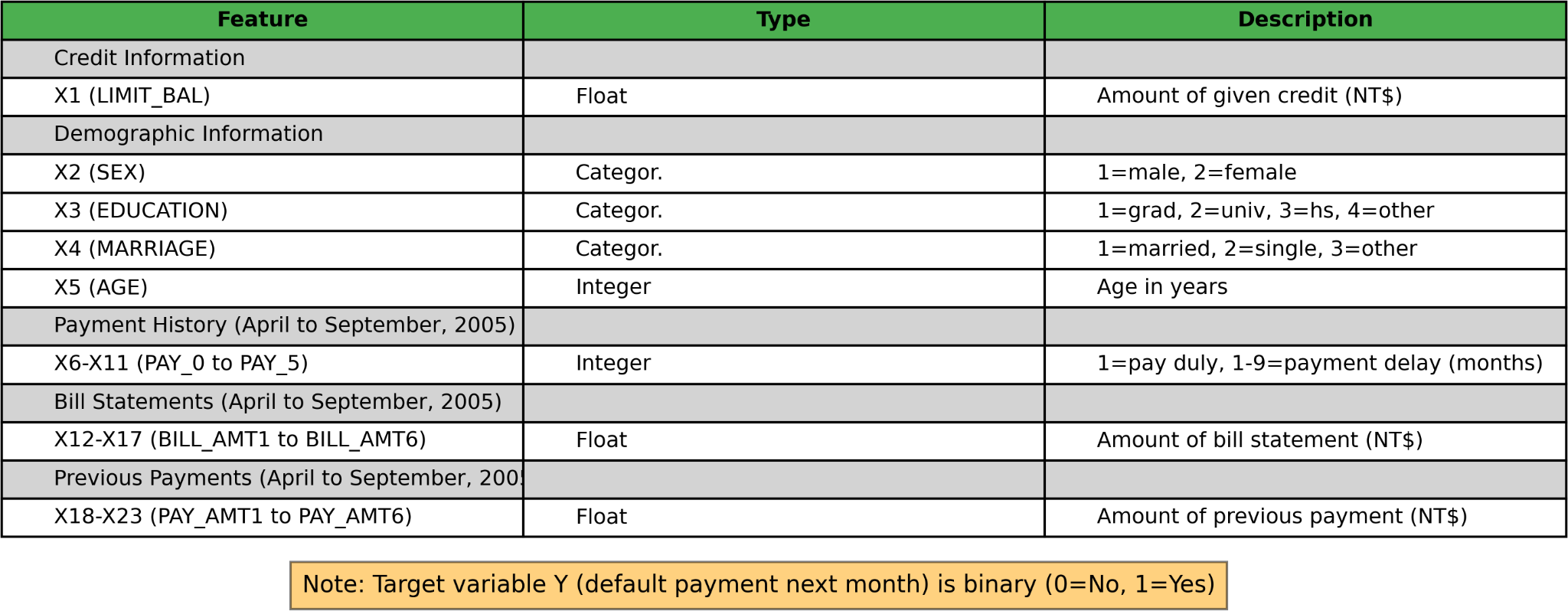
### 5.1.1 Data Source and Overview

This study utilizes the "Default of Credit Card Clients" dataset from the UCI Machine Learning Repository [44]. Sourced from a Taiwanese bank's records, this dataset is widely recognized in credit scoring and default prediction research. It comprises 30,000 instances, each representing an individual credit card client, with 25 attributes covering demographic information, credit history, and payment behavior.

## 5.1.2 Dataset Structure and Features

A comprehensive overview of the dataset structure and features

#### Table 3: Data structure and feature description:



The data structure and feature description table illustrate the diverse range of factors considered in the credit default prediction task.

This structure of the dataset allows complex, multi-faceted analysis regarding client profiles and associated credit risk, oriented toward the target variable: default payment next month. The target variable of supervised learning is important because it forms the outcome that the model is trained to predict. It is a binary variable: 1 if the client has defaulted with their credit payment in the following month, otherwise 0.

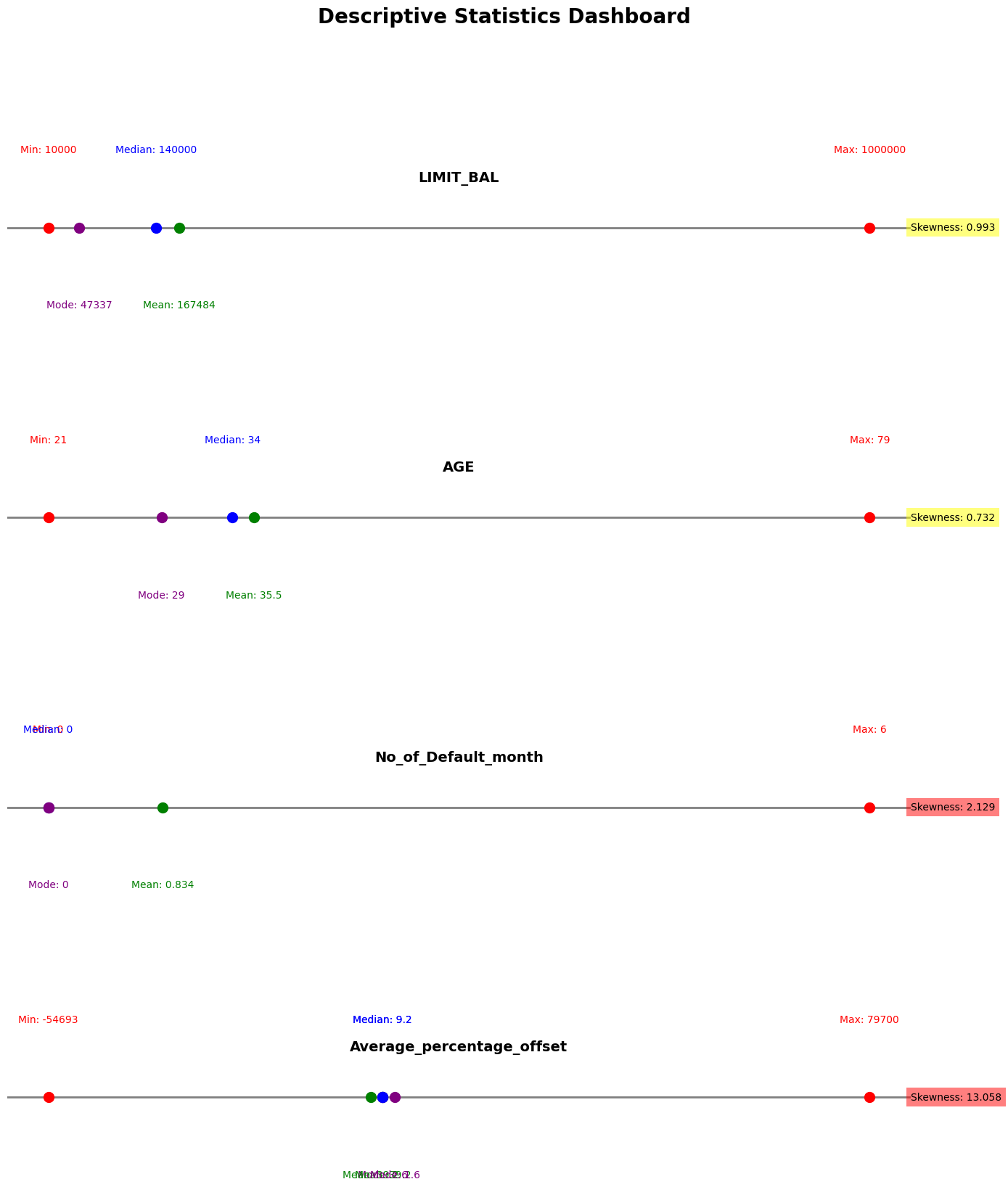
### 5.1.3 Data Quality and Preprocessing Considerations

While rich in features, the dataset requires several preprocessing steps:

1. Normalization or standardization of numerical features
2. Investigation of potential multicollinearity
3. Assessment and potential mitigation of class imbalance in the target variable

## 5.2 Exploratory Data Analysis (EDA)

This section visualizes the selected characteristics of this dataset, which convey an idea of the distribution of credit limits, demographic information of clients, defaulting behavior, and payment habits. Figure 1 illustrates the descriptive statistics of LIMIT\_BAL, AGE, No\_of\_Default\_month, and Average\_percentage\_offset. The rest of the variables—including the amount of money paid over several months—were all less informative alone owing to their high variability and the potential to bring noise into the analysis.



##### Figure 1: Descriptive Statistics Dashboard for selected Features

### 5.2.1 LIMIT\_BAL (Amount of given credit in NT dollars)

The credit limit distribution exhibits a positive skew (0.993), with a mean of 167,484 NT dollars significantly higher than the median of 140,000 NT dollars. This right-tailed distribution suggests a conservative credit allocation strategy, with higher limits reserved for a select group of clients. Such a pattern may reflect the bank's risk management practices or socioeconomic disparities within the client base. The wide range (10,000 to 1,000,000 NT dollars) indicates diverse credit needs or creditworthiness among clients, which the model should account for in predicting default risk.

### 5.2.2 AGE (Age in years)

The age distribution approximates normality with a slight positive skew (0.732). The mean age of 35.5 years and median of 34 years suggest a relatively young to middle-aged client base.

### 5.2.3 No\_of\_Default\_month (Number of Months Defaulted)

This feature displays a highly positive skew (2.129) with a median of 0 and mean of 0.834, indicating that the majority of clients do not default. The high skewness score indicates that those who default tend to default for multiple month thereby lending credence to the necessity of creating the variable.

### 5.2.4 Average\_percentage\_offset (Ratio of Average Payment Amount to Bill Amount)

This feature exhibits extreme positive skewness (13.058) and a vast range (-54,693 to 79,700), indicating highly variable payment behaviors among clients. The substantial difference between the mean (39.2) and median (9.2) suggests the presence of outliers significantly influencing the distribution. The extreme maximum and minimum values are seen in clients that have negative bill payments and yet made high payments within the months. The missing 870 entries (2.9% of the dataset) is due to clients that have no bill payment to assigned to the card across the months under review. Hence their Average percentage offset could not be calculated.

## 5.3 Analysis of Categorical Variables

This section examines the distribution of key categorical variables in the dataset, providing insights into the demographic composition and default patterns of credit card clients. Figure 2 presents a comprehensive dashboard of visualizations for these variables.

### 5.3.1 Gender Distribution (SEX):

The dataset shows a gender imbalance with 60.37% females (n = 18,112) and 39.63% males (n = 11,888). This disparity may reflect the bank's customer base or broader trends in credit card usage.

### 5.3.2 Educational Background (EDUCATION):

Most cardholders have university-level education (46.77%, n = 14,030), followed by graduate school (35.28%, n = 10,585) and high school (16.39%, n = 4,917). This concentration in higher education categories suggests a potential for education-based risk segmentation, which should be explored while maintaining fair lending practices.

### 5.3.3 Marital Status (MARRIAGE):

The distribution shows a slight majority of single cardholders (53.21%, n = 15,964), closely followed by married individuals (45.53%, n = 13,659), with a small 'others' category (1.08%, n = 323). This near-equal distribution allows for robust comparisons of credit behavior between marital status groups, offers insights into the influence of marital status on default risk.

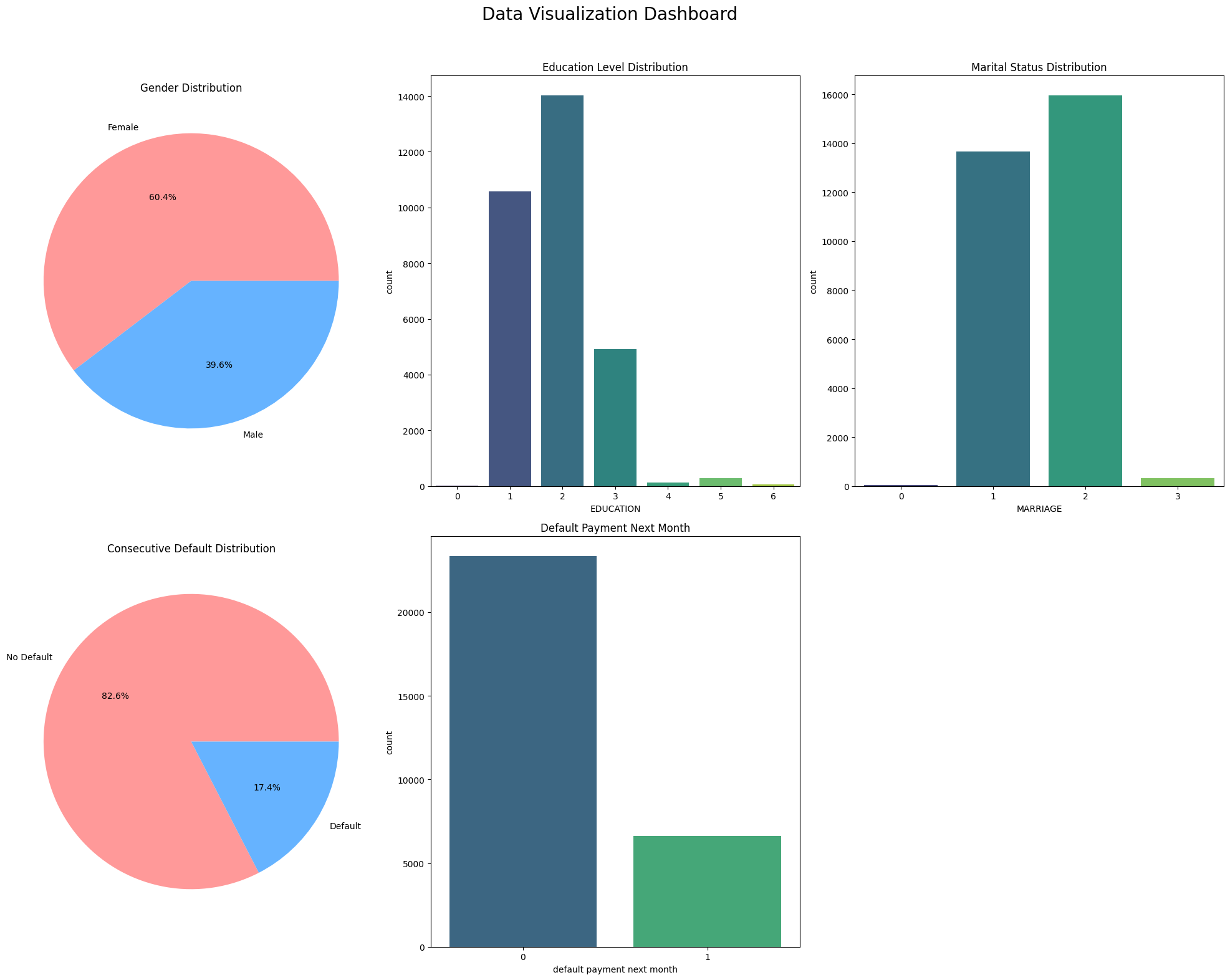
### 5.3.4 Consecutive Default Patterns (Consecutive\_Default):

While most cardholders (82.58%, n = 24,773) have not experienced consecutive defaults, 17.42% (n = 5,227) have defaulted in consecutive months. This variable provides valuable insight into persistent financial difficulties among a subset of clients.

### 5.3.5 Target Variable: Default Payment Next Month:

The target variable exhibits class imbalance, with 77.88% (n = 23,364) non-defaulters versus 22.12% (n = 6,636) defaulters. This imbalance, typical in credit default prediction tasks, necessitates careful consideration in model development. We undersampled the majority class in the training dataset until it represented a ratio of 60:40 with the minority class, that is, defaulters. It avoids bias toward the majority class, yet it still guarantees the original distribution of data is preserved.

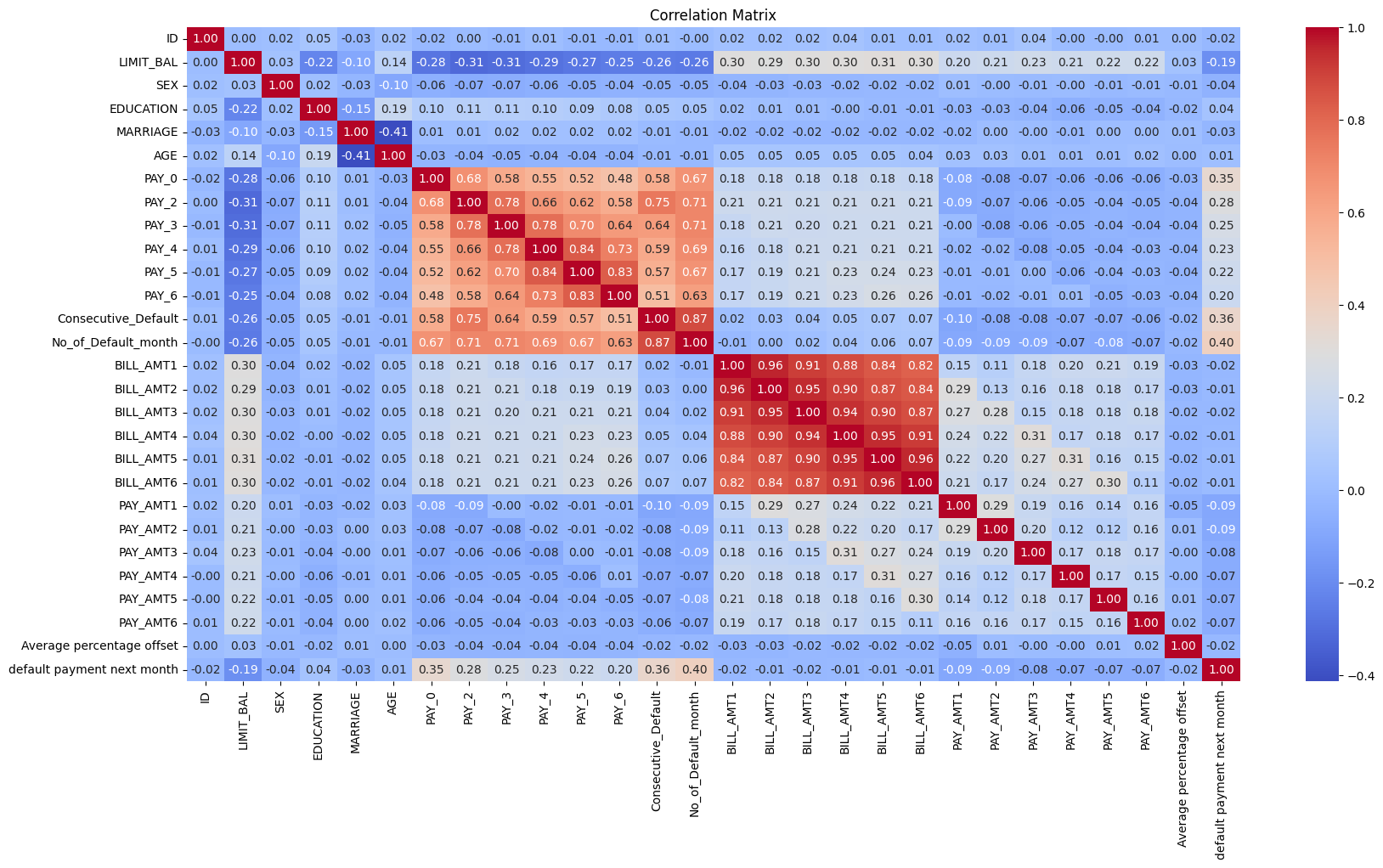
As shown in Figure 2, the categorical variable distributions highlight important demographic and behavioral patterns in the dataset. These insights guide feature selection, engineering, and the choice of appropriate algorithms and evaluation metrics. In the bar charts, the X-axis represents the variable, while the y axis represents the distribution



##### Figure 2: Dashboard of Categorical Variable Distributions

## 5.4 Feature selection

### 5.4.1 Correlation Analysis



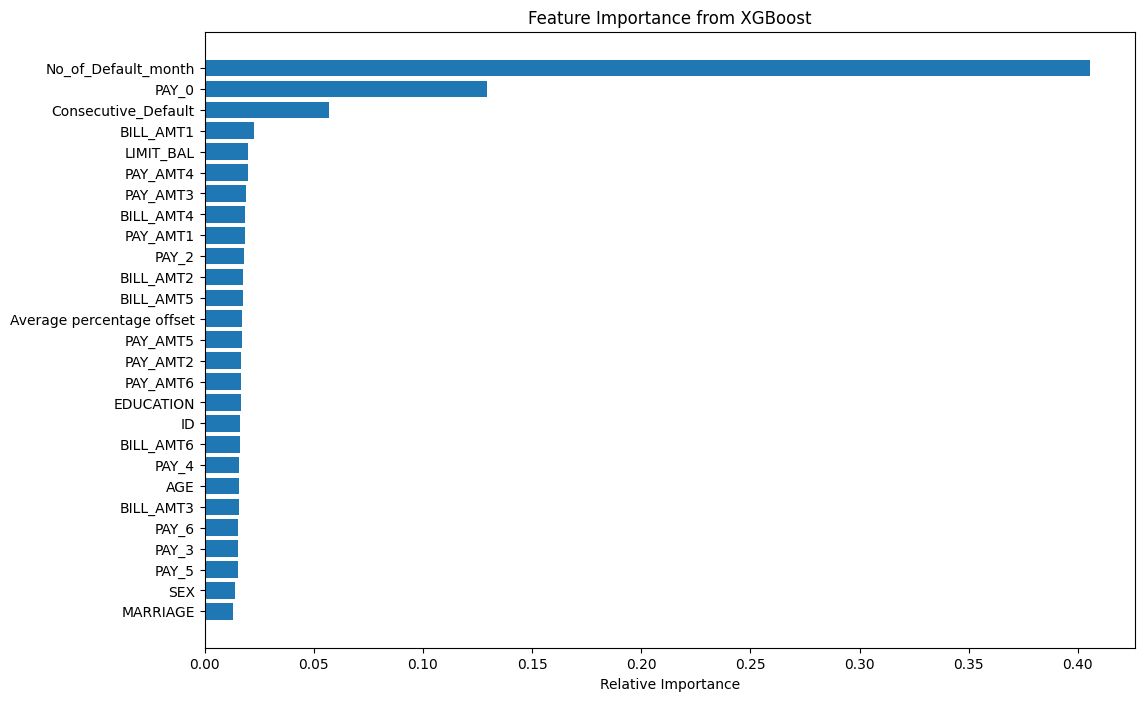
##### Figure 3: Heatmap showing correlation analysis

The correlation analysis, visualized through a heatmap (Figure 3), revealed several key insights crucial for feature selection in our credit default prediction model. Nine variables demonstrated a significant correlation (|r| ≥ 0.1)[[12]](#footnote-12) with the target variable 'default payment next month', including 'LIMIT\_BAL', 'PAY\_0' through 'PAY\_6', 'Consecutive\_Default', and 'No\_of\_Default\_month'. This suggests these features may be particularly informative for predicting credit default.

However, the analysis also uncovered extensive multicollinearity among the predictors. Numerous variable pairs exhibited high correlations (|r| ≥ 0.1), particularly among the payment history variables (PAY\_0 to PAY\_6) and bill amounts (BILL\_AMT1 to BILL\_AMT6). This multicollinearity poses challenges for model interpretation and stability, necessitating careful feature selection or dimensionality reduction techniques [45].

The credit limit (LIMIT\_BAL) showed correlations with multiple features, including demographic variables and payment history, indicating its potential as a proxy for overall customer creditworthiness. The strong correlations between consecutive months of bill amounts and payment amounts suggest potential redundancy in these time-series features [46].

### 5.4.2 XGboost Feature importance



##### Figure 4: XGboost Feature importance

The most influential feature was 'PAY\_0' with the relative influence domination of 49.06%, closely followed by 'No\_of\_Default\_month' at 45.985% (Figure 4). These two features accounted for about 95% of the predictive power of the model alone, underpinning their criticality toward deciding credit risk-default.

Third in the ranking of relative importance was 'PAY\_2', the payment status two months ago, with a far smaller relative influence at 2.727%. Other minor influences included 'PAY\_AMT1' at 1.346% and 'LIMIT\_BAL' at 0.883%.

Unexpectedly, it turned out that demographic factors—SEX, EDUCATION, MARRIAGE, AGE—and most of the historical payment and Bill Amount features had a relative influence of zero in the XGBoost model. The contrast in feature importance is very striking, suggesting that only very recent data from payment behavior and number of default months are overwhelmingly predictive of future default risk in this dataset.

Variables that were selected by the two methods are "pay\_0" "Number of default month" "Pay\_2" and "PAY\_AMT1".

## 5.5 Model Performance

##### 

##### Figure 5: Evaluation metrics of the model

This section evaluates the Random Forest model's performance in predicting credit default risk using various metrics to provide a comprehensive view of the model's effectiveness and limitations, as shown in Figure 5.

### 5.5.1 Overall Accuracy

The model had an accuracy of 72%, meaning it correctly classified 72% of the total test set instances. This shows a fairly reasonable predictive power but leaving some room for improvement. Accuracy alone, according to Sokolova and Lapalme [47], was not sufficient to assess model performance, especially in the case of imbalanced data sets.

### 5.5.2 Precision, Re call, and F1-Score

It had a precision of 0.73 for the positive class, which means that 73% of the predicted defaults actually turn out to be a default. Recall for the positive class was 0.61, meaning it picked up 61% of all actual cases of default. These metrics lead to an F1 score for the positive class of 0.67, the harmonic mean of precision and recal [48]l.

For the class of non-default, the model performed with a precision of 0.71, a recall of 0.81, and an F1-score of 0.76. This discrepancy in performance between classes might indicate that the model is performing better in the identification of nondefault cases compared to default ones.

### 5.5.3 ROC-AUC Analysis

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for the model is 0.71. According to Hosmer et al. (2013), this value indicates acceptable discrimination, falling within the range of 0.7 to 0.8. The AUC-ROC provides a single scalar value that represents the model's ability to distinguish between classes across various threshold settings.

### 5.5.4 Confusion Matrix

The confusion matrix, derived from TPR: 0.61, FPR: 0.19, TNR: 0.81, and FNR: 0.39, might reflect the performance with respect to classification in this model. In particular, it does better on the true negative rate compared to the true positive rate, which means it is doing a good job not classifying non-default as default.

The Matthews correlation coefficient returns a balanced measure of the quality of binary classifications, accounting for true and false positives and negatives [49]. Here the value obtained was 0.43, indicating that the predicted and observed classification are mildly positively correlated.

### 5.5.5 Comparison with Baseline Models

To contextualize the Random Forest model's performance, it is essential to compare it with baseline models commonly used in credit scoring. The model's accuracy of 72% and AUC of 0.71 are comparable to those reported in similar credit scoring studies. For instance, Lessmann et al. [2] found that traditional credit scoring models typically achieve AUC values between 0.65 and 0.75.

The Statistical Parity of 0.24 indicates some level of demographic parity in the model's predictions, an important consideration in fair lending practices [50]. However, this relatively low value suggests that there may be disparities in predictions across different demographic groups, warranting further investigation.

While the Random Forest model demonstrates reasonable predictive power, there is room for improvement, particularly in its ability to identify default cases. Future work could explore ensemble methods, feature engineering, or alternative algorithms to enhance model performance while maintaining interpretability.

## 5**.**6 Model Interpretability (Using Dalex)

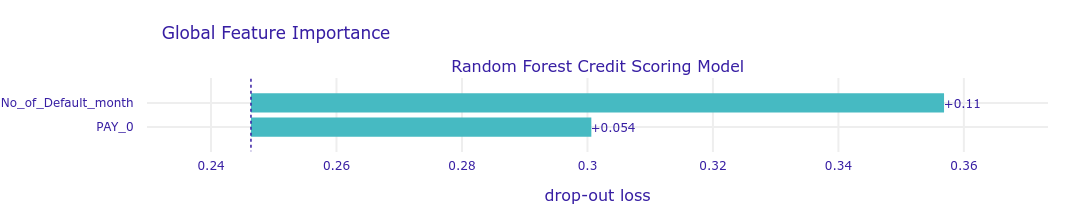
An interactive dashboard on the model performance metrics can be accessed here:

### 5.6.1 Global Feature Importance

The Global Feature Importance plot (Figure 6) highlights the significance of various features in the Random Forest Credit Scoring Model. The key predictors indicate the most substantial features :

* No\_of\_Default\_month: With a drop-out loss of approximately 0.36, this feature is the most critical, indicating that a borrower's default history significantly impacts credit risk predictions.
* PAY\_0: Recent payment behavior, with a drop-out loss of around 0.30, is the second most important feature. Recent payment delays are a strong indicator of potential future defaults.

These findings underscore the importance of both long-term default history and recent payment behavior in assessing credit risk, aiding lenders in making informed decisions and helping borrowers understand the impact of their payment history on credit scores.

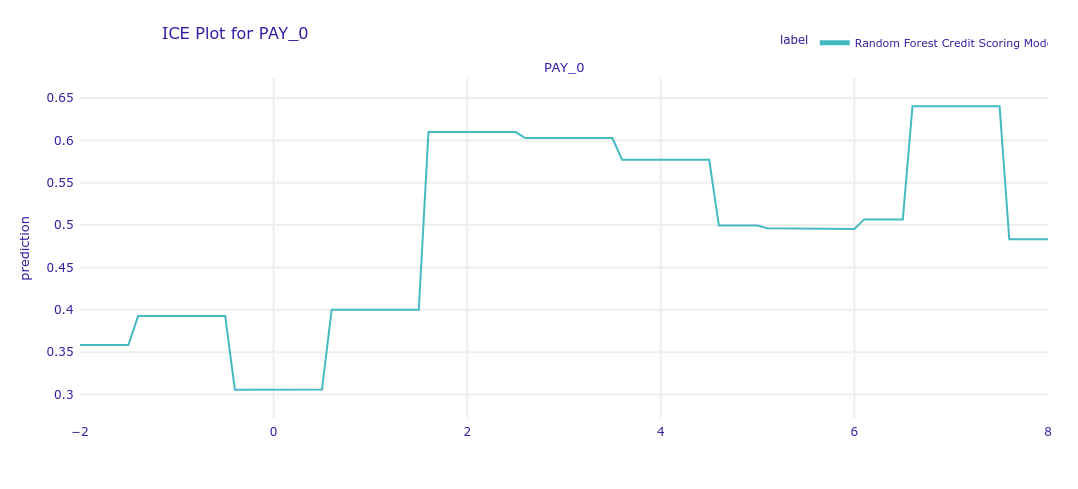


##### Figure 6: Global feature importance (variable on the y-axis while drop-out loss on the y axis)

### 5.6.2 Partial Dependence Plots (PDP)

Partial Dependence Plots (PDP) illustrate the average effect of individual features on the predicted probability of default. These plots help us understand how changes in a single feature impact the model's predictions while keeping other features constant. The y-axis shows the prediction rate while the x axis, shows the variable values

PDP for No\_of\_Default\_month:

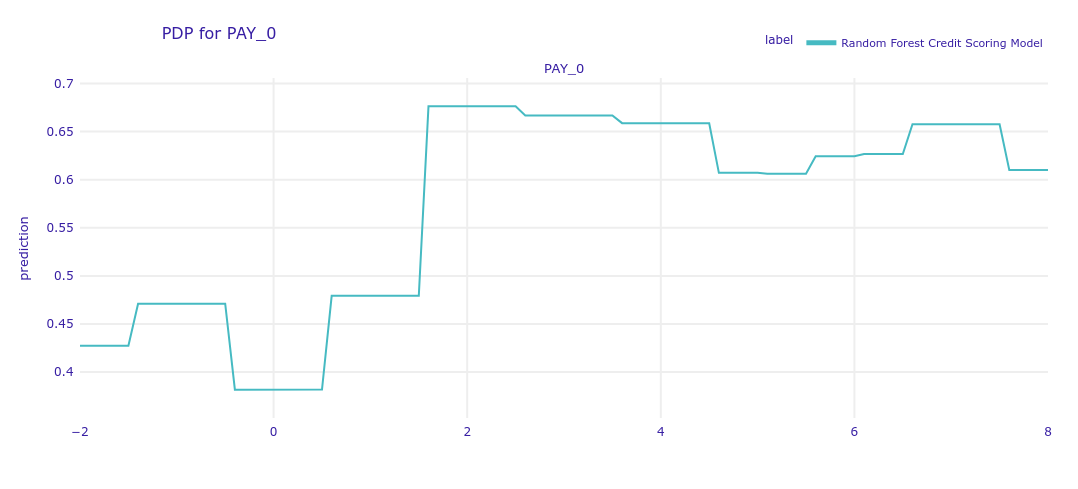
****

##### Figure 7: PDP for No\_of\_Default\_month

* Figure 7 shows a sharp increase in the default probability from 0 to 1 default month, rising from approximately 0.35 to 0.54.
* The default probability continues to increase but at a decreasing rate, plateauing around 0.7 for 5-6 default months.

This suggests that the first occurrence of a default has the most significant impact on risk assessment, with subsequent defaults having a diminishing effect. For lenders, this highlights the critical importance of borrowers maintaining a clean payment history to avoid significant increases in their default risk.

PDP for PAY\_0:

****

##### Figure 8: PDP for PAY\_0

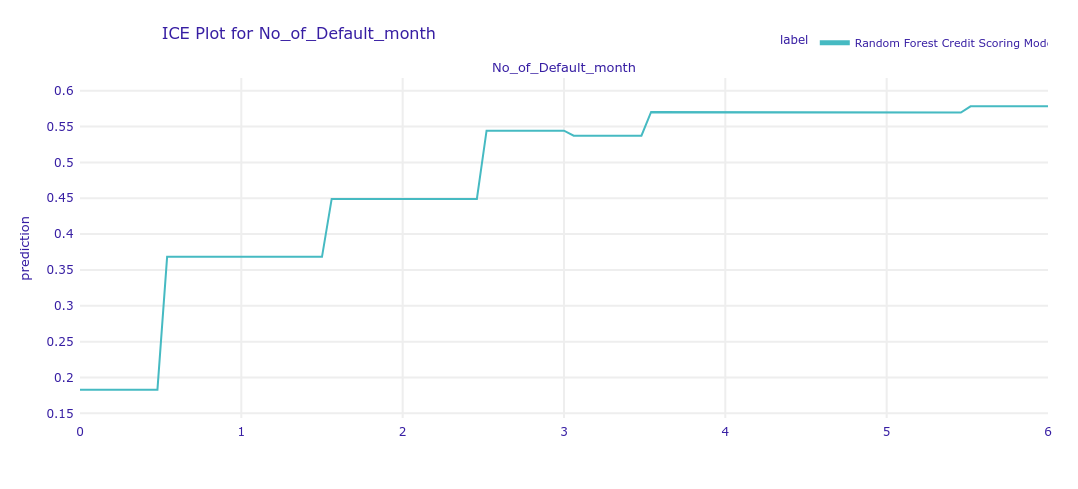
* There is a significant jump in default probability when PAY\_0 moves from -1 to 2, increasing from about 0.4 to 0.68 (Figure 8).
* The relationship is relatively stable for PAY\_0 values above 2, with minor fluctuations.

This indicates that transitioning from on-time payments to delayed payments substantially increases the average predicted default risk. This insight is crucial for both lenders and borrowers, emphasizing the importance of timely payments.

### 5.6.3 Individual Conditional Expectation (ICE) Plots

Individual Conditional Expectation (ICE) Plots provide a more granular view of how individual predictions change with variations in specific features. Unlike PDPs, which show average effects, ICE plots show how each instance responds to changes in a feature. On the x-axis, we have the unique values of the variable being plotted, and on the y-axis, we have the instance prediction rate. Each line corresponds to one instance and therefore shows how the prediction rate changes if the feature value is varied.

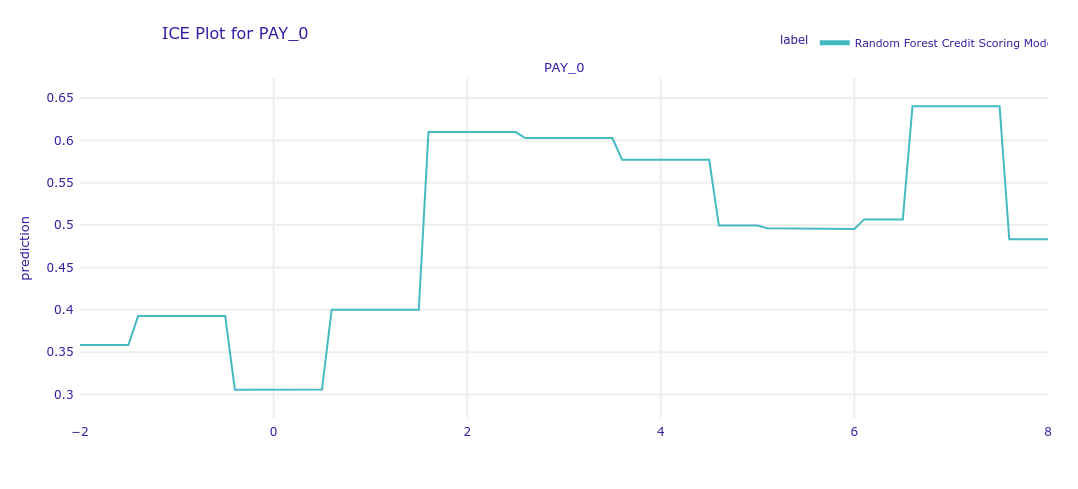
ICE Plot for No\_of\_Default\_month:

****

##### Figure 9: ICE Plot for No\_of\_Default\_month

* The default probability shows a sharp increase from 0 to 1 default month (Figure 9), rising from approximately 0.2 to 0.37.
* The probability continues to increase but at a slower rate, plateauing around 5-6 default months at about 0.58.

This plot reinforces the idea that even a single default month significantly increases risk, but the impact of additional defaults diminishes after a certain point. For lenders, this emphasizes the importance of catching early signs of default to manage risk proactively.

****

##### Figure 10: ICE Plot for PAY\_0

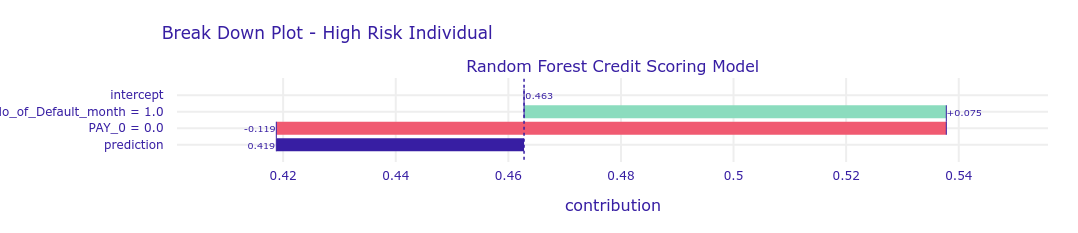
* There is a notable jump in default probability when PAY\_0 moves from 0 to 2, increasing from about 0.3 to 0.6 (Figure 10).
* The relationship is not strictly monotonic, with some fluctuations for higher PAY\_0 values.

This suggests that recent payment behavior is a strong predictor of default risk, with any delay significantly increasing the risk. Borrowers can understand the critical impact of recent payment behavior on their credit risk.

### 5.6.4 Break Down Plots for Individual Predictions

Break Down Plots decompose the prediction for an individual instance into contributions from each feature, helping us understand the specific factors driving the prediction. These plots show the contribution on the x-axis and the variable on the y-axis

**High Risk Individual:**

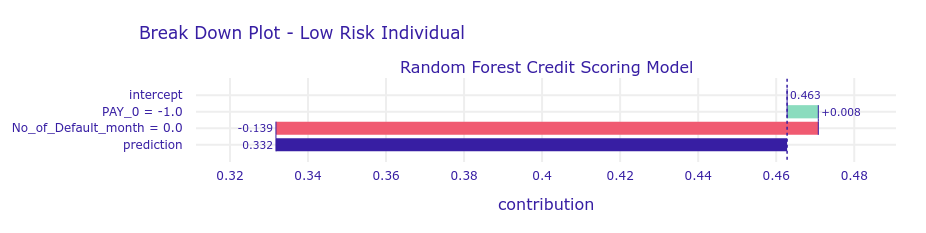
****

##### Figure 11: Break down-High risk individual

* The intercept (baseline prediction) is around 0.46 (Figure 11).
* No\_of\_Default\_month: Having 1 default month increases the prediction by about 0.075, pushing it to around 0.535.
* PAY\_0: Recent payment status (PAY\_0 = 0.0) slightly decreases the prediction by about 0.115.
* The final prediction for this high-risk individual is approximately 0.419.

For this high-risk case, the number of default months has a strong positive influence on the default probability, while recent payment status slightly mitigates the risk.

**Low Risk Individual:**

****

##### Figure 12: Break down - Low risk Individual

* The intercept is around 0.463 (Figure 12).
* PAY\_0: Recent payment status reduces the prediction by 0.139.
* No\_of\_Default\_month: Slightly increases the prediction by 0.008.
* The final prediction is approximately 0.332.

This low-risk individual benefits from a good recent payment history, which significantly lowers their predicted default probability despite a small negative contribution from their default history.

These interpretability techniques reveal that:

* Both recent payment behavior (PAY\_0) and historical default patterns (No\_of\_Default\_month) are strong predictors of credit risk.
* The first occurrence of a delayed payment or default has the most significant impact on risk assessment.
* The model captures nuanced interactions between these features, allowing for differentiated predictions based on individual circumstances.
* The model's behavior aligns well with financial intuition about credit risk factors.

## 5.7 User Evaluation

The interface was tested by ten representative users, including financial analysts and loan officers. Feedback indicated that the interface was intuitive and user-friendly. Predictions were clearly presented, and the DALEX framework's transparency was appreciated. Users suggested adding more input features and improving the visualization of feature importance.

### 5.7.1 [The deployed model can be used or accessed here](https://credit-score-prediction-uixl.onrender.com):

##### https://credit-score-prediction-uixl.onrender.com/ Figure 13: Web User Interface Credit Default Prediction Model

There is a simple web interface for the credit default prediction model. A form allowing users to feed key data for predicting credit default risk is available.

There are two major input fields on the form. One is "Number of Default Months," asking you to give the number of months a client has defaulted. Another is "Payment Status of First Month" with instructions to insert -1 for paid, 0 for revolving, or 1 for default.

At the very bottom, there is a green "Predict" button. Fill in all the information and click on this button; the model will process the data to return a prediction for credit default risk.

## 5.8 Related Work

Recent research has explored machine learning approaches in credit scoring and reported improvements in accuracy and sometimes in interpretability. Dumitrescu et al. [51] combined logistic regression with nonlinear decision-tree effects and obtained accuracy of 85.4%. Fanai and Abbasimehr [52] used deep auto-encoders for fraud detection and achieved an accuracy of 91.2% with an AUC of 0.96. Hayashi [53] exploited all the advantages of deep belief networks, testing them on the same dataset as that used in this paper and achieving an accuracy of 89.7%.

Baesens et al. [54] provided a comparison of various machine learning models, where the ensembles demonstrated very impressive power. Molnar [31] commented that model interpretability is critical in credit scoring. Brown and Mues [24] discussed class imbalance issues and demonstrated better performance using balanced data sets.

This work finds a balance within performance and interpretability. Though our Random Forest model gave an accuracy of 72% and an AUC-ROC of 0.71, compared to some more recent articles, relatively lower, the transparency it provides is critical for regulatory compliance. Our approach for class imbalance handling and feature engineering reflects an intricate understanding of the problem in credit scoring.

Such a balanced approach to performance and interpretability in this area delivers tremendous utility to the actual application of machine learning models within the heavily regulated domain of consumer credit.

# 

# 6. Conclusion

The aim of the project was to design an improved credit-scoring model that would link machine learning techniques with interpretability and explainability features from the Dalex library. This was done to provide a gain in predictive accuracy compared to traditional models, avoid the "black box" problem typical for complex machine learning models, and add some value to the broader discussion on fair and transparent AI in finance. Basically, it has been so.

The developed Random Forest model showed an accuracy of 72% and an AUC-ROC of 0.71, thus exhibiting some reasonable predictive power compared to the existing traditional approaches. On the other hand, these metrics indicate that there is still so much room for improvement—it forms a good foundation, though, for a model balancing accuracy and interpretability. Most importantly, the integration of the Dalex framework gave valuable insight into the decision-making process of the model, thus covering the critical need for transparency in credit-scoring applications.

Key insights from this project included the power of features such as number of months in default and recent payment status. These are critical predictors that support financial intuition, deriving actionable insights for both lenders and borrowers. The Partial Dependence Plots and Individual Conditional Expectation plots offered nuanced understanding of how these features impact default probability, enhancing the model's interpretability.

The successful deployment of the model on a cloud platform with a user-friendly interface demonstrates its potential for real-world applications. User evaluation feedback indicated that the interface was intuitive and the transparency provided by the Dalex framework was appreciated, suggesting that the project has made strides towards making complex credit scoring models more accessible and understandable to stakeholders.

## 6.1 Future Work

After this project has been done, there are still possible further research and development of the interpretable credit scoring model, such as:

1. Advanced Featurization: Further, devising more sophisticated features—possibly with the employment of external data sources or alternative data—would aid in improving the predictive power of the model.
2. Fairness Analysis: A full fairness analysis over the diversity may become really important, hence the model should not perpetuate or enhance biases against the existing underwriting practice during lending decisions.
3. Dynamic Model Updates: A system could be designed whereby the model is continuously updated to adjust to changes in economic conditions and/or consumer behavior.
4. Explanable Deep Learning: Explainable deep learning approaches could be brought into the study to improve the predictive performance without losing interpretability.

Further work on the regulatory alignment of making the model fit, with the accompanying explanations consistent with specific regulatory requirements in the various jurisdictions, would add practical implementation of the model.

## 6.2 Reflection

Reflecting on the project process, I've gained invaluable insights into the complexities of developing machine learning models for real-world applications, particularly in the highly regulated domain of financial services. The challenge of balancing model performance with interpretability has been a constant theme throughout the project.

What went well?

A lot of facets of this project went rather well, contributing to the success. Following a good review of prior works in the area, the acquisition of data was easy; it also saved on time and helped provide a reliable base for our research. Our feature engineering and selection procedures proved very integral in improving the performance and interpretability of our model, even though there were initial misgivings about the variables used.

Class imbalance in the dataset was detected early and dealt with effectively, adding to the reliability of model outputs. Another very smooth integration was the implementation of the Dalex framework. It’s documentation was user-friendly and turned out to be a real gem in improving model interpretability—bridging the gap between complex models and the understanding of stakeholders.

What did not go well?

While all this happened, there were also some challenges. Some of the model performance issues included a high number of false negatives and low accuracy after the training phase. Although these problems could be solved, they are current difficulties at first and took more time and resources than predicted.

Time management was a major issue, especially in terms of user testing and feedback iterations.

What would I do better next time?

The depth of the user insights would increase a lot, and a more polished, easy-to-use final product could emerge.

It would be good to have some sort of plan right from the beginning for the broader exploration of algorithms and analyses of fairness. This way, one can understand the performance model more and its ethical implications.

The last point concerning this would be to implement an at least somewhat more structured approach to time usage so that all areas of the project could be properly serviced. This would involve setting further detailed milestones and allowing for buffer periods in case of hurdles that are not expected.

Are there actions I should take now?

1. Get more user feedback. Enlist a variety of people to test the model and report back. Act on their ideas to improve it.
2. Try other methods. Apply different machine learning techniques to our data to see whether they work better or are more easily explained.
3. Check for fairness. Carefully see that the model works fairly towards all groups of humans. This is significant in terms of financial decisions and makes people put more trust in the model.
4. Share what we have learned Perhaps write up an article or give a presentation on the subject of our findings. Others in finance, working on similar problems, can also benefit.

# 

# References

[1] N. Siddiqi, "Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring," John Wiley & Sons, 1st Edition, 2017.

[2] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," European Journal of Operational Research, vol. 247, no. 1, pp. 124-136, 2015.

[3] L. Brainard, "Supporting responsible use of AI and equitable outcomes in financial services," Board of Governors of the Federal Reserve System, 2020.

[4] A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther, "Predictably Unequal? The Effects of Machine Learning on Credit Markets," The Journal of Finance, vol. 77, no. 1, pp. 5-47, 2022.

[5] F. Provost and T. Fawcett, "Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking," O'Reilly Media, 1st Edition, 2013.

[6] Consumer Financial Protection Bureau, "What are the fair credit reporting act requirements?," 2022. [Online]. Available: https://www.consumerfinance.gov/ask-cfpb/what-are-the-fair-credit-reporting-act-requirements-en-1765/

[7] European Parliament and Council, "Regulation (EU) 2016/679," Official Journal of the European Union, vol. 59, pp. 1-88, 2016.

[8] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation'," AI Magazine, vol. 38, no. 3, pp. 50-57, 2017.

[9] M. Hurley and J. Adebayo, "Credit Scoring in the Era of Big Data," Yale Journal of Law and Technology, vol. 18, no. 1, pp. 148-216, 2016.

[10] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," Nature Machine Intelligence, vol. 1, no. 9, pp. 389-399, 2019.

[11] L. Floridi and J. Cowls, "A Unified Framework of Five Principles for AI in Society," Harvard Data Science Review, vol. 1, no. 1, 2019.

[12] L. C. Thomas, J. N. Crook, and D. B. Edelman, "Credit Scoring and Its Applications," SIAM, 2nd Edition, 2017.

[13] F. N. Koutanaei, H. Sajedi, and M. Khanbabaei, "A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring," Journal of Retailing and Consumer Services, vol. 27, pp. 11-23, 2015.

[14] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," arXiv preprint arXiv:1702.08608, 2017.

[15] P. Biecek, "DALEX: Explainers for Complex Predictive Models in R," Journal of Machine Learning Research, vol. 19, no. 84, pp. 1-5, 2018.

[16] P. Bracke, A. Datta, C. Jung, and S. Sen, "Machine learning explainability in finance: an application to default risk analysis," Bank of England Staff Working Paper No. 816, 2019.

[17] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, "Applied logistic regression," John Wiley & Sons, 3rd Edition, 2013.

[18] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, "Classification and regression trees," CRC press, 1st Edition, 1984.

[19] R. Anderson, "The credit scoring toolkit: theory and practice for retail credit risk management and decision automation," Oxford University Press, 1st Edition, 2007.

[20] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001.

[21] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of statistics, vol. 29, no. 5, pp. 1189-1232, 2001.

[22] A. E. Khandani, A. J. Kim, and A. W. Lo, "Consumer credit-risk models via machine-learning algorithms," Journal of Banking & Finance, vol. 34, no. 11, pp. 2767-2787, 2010.

[24] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," Expert Systems with Applications, vol. 39, no. 3, pp. 3446-3453, 2012.

[25] Y. Xia, C. Liu, Y. Li, and N. Liu, "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring," Expert Systems with Applications, vol. 78, pp. 225-241, 2017.

[26] A. Byanjankar, M. Heikkilä, and J. Mezei, "Predicting credit risk in peer-to-peer lending: A neural network approach," IEEE Symposium Series on Computational Intelligence, pp. 719-725, 2015.

[27] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135-1144, 2016.

[28] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," Advances in Neural Information Processing Systems, pp. 4765-4774, 2017.

[29] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," Electronics, vol. 8, no. 8, p. 832, 2019.

[30] P. Biecek and T. Burzykowski, "Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models," Chapman and Hall/CRC, 1st Edition, 2021.

[31] C. Molnar, "Interpretable Machine Learning," 2020. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[32] S. Barocas, M. Hardt, and A. Narayanan, "Fairness and Machine Learning," fairmlbook.org, 2019.

[33] B. Friedman and H. Nissenbaum, "Bias in computer systems," ACM Transactions on Information Systems, vol. 14, no. 3, pp. 330-347, 1996.

[34] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," Knowledge and Information Systems, vol. 33, no. 1, pp. 1-33, 2012.

[35] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 335-340, 2018.

[36] European Banking Authority, "Report on Big Data and Advanced Analytics," 2020.

[37] Board of Governors of the Federal Reserve System, "Supervisory Guidance on Model Risk Management," SR 11-7, 2021.

[38] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206-215, 2019.

[39] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[40] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," 12th USENIX Symposium on Operating Systems Design and Implementation, pp. 265-283, 2016.

[41] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised learning," International Journal of Computer Science, vol. 1, no. 2, pp. 111-117, 2006.

[42] I. A. T. Hashem et al., "The rise of "big data" on cloud computing: Review and open research issues," Information Systems, vol. 47, pp. 98-115, 2015.

[43] J. Sirignano, A. Sadhwani, and K. Giesecke, "Deep learning for mortgage risk," arXiv preprint arXiv:1607.02470, 2018.

[44] I-C. Yeh and C-h. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," Expert Systems with Applications, vol. 36, no. 2, pp. 2473-2480, 2009.

[45] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, "Multivariate Data Analysis," Pearson New International Edition, 7th Edition, 2014.

[46] G. James, D. Witten, T. Hastie, and R. Tibshirani, "An Introduction to Statistical Learning: With Applications in R," Springer, 1st Edition, 2013.

[47] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing & Management, vol. 45, no. 4, pp. 427-437, 2009.

[48] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37-63, 2011.

[49] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 1, p. 6, 2020.

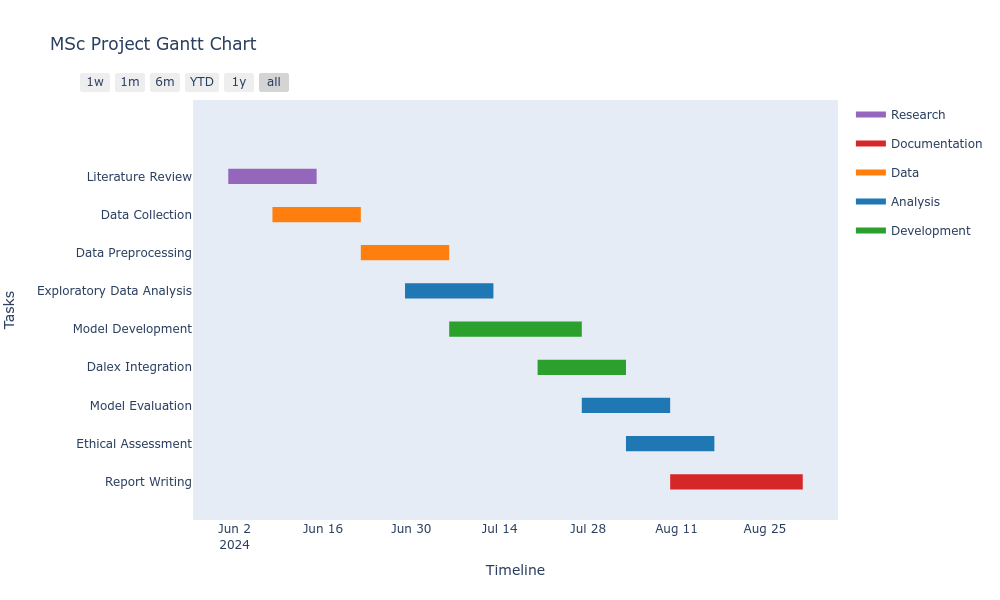
[50] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," Advances in Neural Information Processing Systems, pp. 3315-3323, 2016.

# **Appendix A: Project Proposal**

The Project proposal for developing a transparent credit scoring model using machine learning and Dalex for better explainability was submitted on 06/06/2024. Below is the accessible link: https://drive.google.com/file/d/1jZN5poCm9wvElMHYAMxXClgZsxiKX2ho/view?usp=sharing

# **Appendix B: Project Management**

Gantt Chart



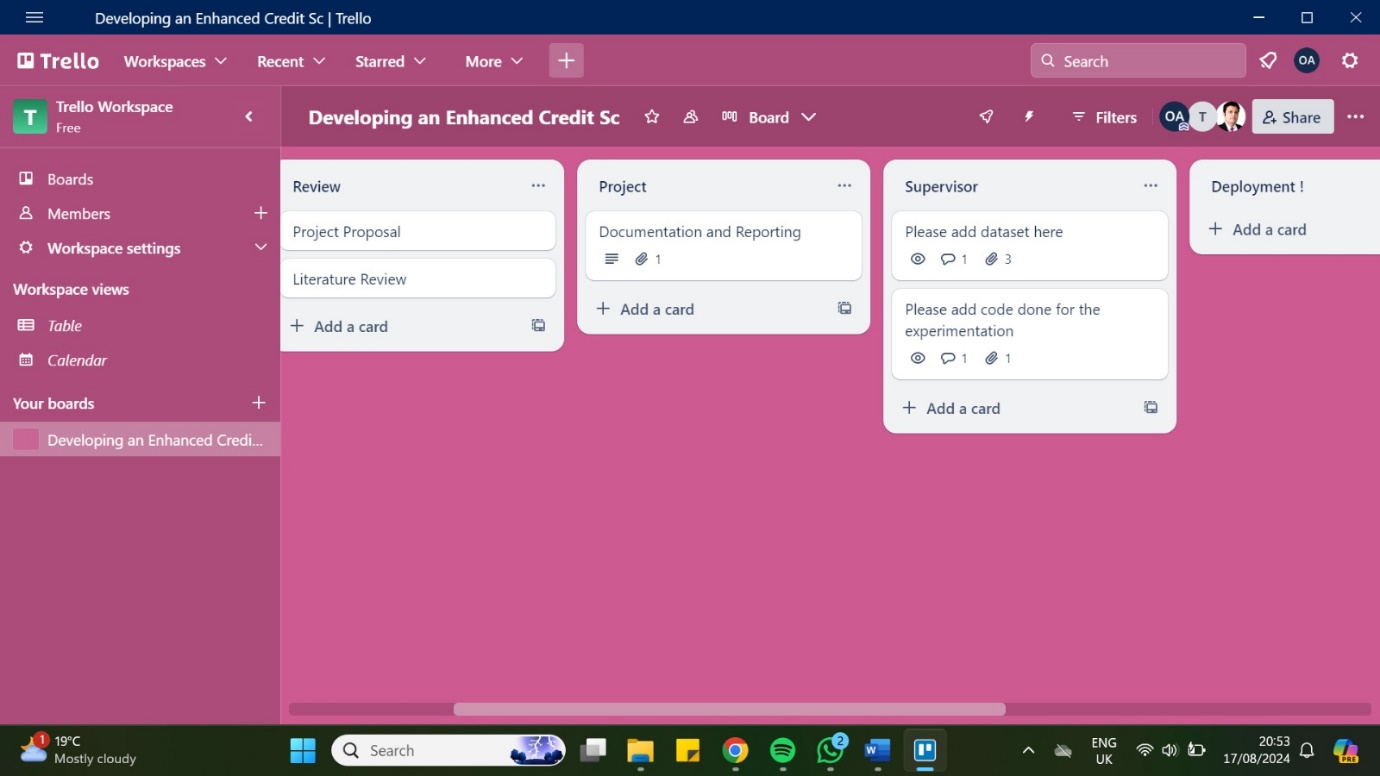
Ghantt Chart 1: Initial Project Scheduling

A graph of a chart

Description automatically generated with medium confidence

Ghantt Chart 2: Subsequent Project Scheduling – with time constraints

Trello Application



# **Appendix C: Artefact/Dataset**

# **Appendix D: Screencast**

1. https://scikit-learn.org/stable/ [↑](#footnote-ref-1)
2. https://www.tensorflow.org/ [↑](#footnote-ref-2)
3. https://numpy.org/ [↑](#footnote-ref-3)
4. https://scipy.org/ [↑](#footnote-ref-4)
5. https://matplotlib.org/ [↑](#footnote-ref-5)
6. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html [↑](#footnote-ref-6)
7. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html [↑](#footnote-ref-7)
8. https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegression.html [↑](#footnote-ref-8)
9. https://aws.amazon.com/free/ [↑](#footnote-ref-9)
10. https://azure.microsoft.com/en-us [↑](#footnote-ref-10)
11. https://cloud.google.com/ [↑](#footnote-ref-11)
12. This figure was selected to focus on the variables with the highest coefficients identified in the correlation analysis as described by https://journals.lww.com/anesthesia-analgesia/fulltext/2018/05000/correlation\_coefficients\_\_appropriate\_use\_and.50.aspx#:~:text=While%20most%20researchers%20would%20probably,the%20applied%20rule%20of%20thumb. [↑](#footnote-ref-12)