National College of Ireland

Configuration Manual

MSc Research Project
MSc. Financial Technology

Emmanuel Mani
Student ID: 22211535

School of Computing
National College of Ireland

Supervisor: Brian Byrne

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……Emmanuel Mani……

Student ID: ……22211535……..……

Programme: ……MSc. Financial Technology…… **Year:** …2023 – 24…..

Module: ……..……MSc Research Programme……..……..……..……

Lecturer: ……Brian Byrne……..……..……

Submission Due Date: ……12-08-2024……..……..……..……..……

Project Title: ……Mahine Learning Frontiers in FinTech: Transforming ……..Credit Risk Assessment……..……..……..……

Word Count: ……5…… **Page Count:** ……885……..……..……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Emmanuel Mani……

Date: ……12-08-2024……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Forename Surname
Student ID:

1 Section 1: Project Overview

This section provides an overview of the project, including its objectives, significance, and a brief description of the models and techniques used.

1.1 Introduction

The primary objective of the project is to evaluate credit risk using machine learning models. Credit risk can be explained as the likelihood that a borrower will default on a loan, which is probably one of the significant problems confronted by any financial institution. Credible predictions of credit risk allow a lender to make wise decisions and minimize losses in money matters.

1.2 Objectives

The main objectives of this research are:

- Performance evaluation of various machine learning models in credit risk prediction.
- Performance comparisons for Logistic Regression, Random Forest, Gradient Boosting Machine, and an Ensemble Model.
- To identify key features that contribute to accurate credit risk prediction.

1.3 Project Significance

This project is significant for the FinTech industry as it explores advanced methods to enhance credit risk assessment, potentially leading to more reliable and efficient lending practices.

2 Section 2 : System Configuration

This section describes the software, libraries, and hardware configurations used in the project.

2.1 Software and Libraries

The following software and libraries were used in this project:

- **Python**: The primary programming language used for data processing, model development, and evaluation.
- **Google Colab**: Used as the development environment due to its accessibility and computational resources.

- **Libraries**:
    - *pandas*: For data manipulation and preprocessing.
    - *numpy*: For numerical computations.
    - *scikit-learn*: For implementing machine learning models like Logistic Regression, Random Forest, and the Ensemble Model.
    - *XGBoost*: For implementing the Gradient Boosting Machine (GBM) model.
    - *matplotlib and seaborn*: For data visualization and generating plots.

2.2 Hardware Configuration

- **Processor**: Google Colab's default hardware configuration with access to GPUs.
- **Memory**: 12 GB RAM provided by Google Colab.

3 Section 3 : Model Implementation

This section details the steps taken to implement the machine learning models.

3.1 Data Preprocessing

- **Handling Missing Values**: The median imputation for numerical variables and the mode for categorical variables are conducted wherever there are missing values.
- **Encoding Categorical Variables**: These variables were categorical, and hence one-hot encoded to put them into a numerical format.
- **Outlier Management**: The outliers were capped at the 99th percentile to reduce their contribution in model predictions.
- **Feature Engineering:** Interaction terms, temporal features and aggregated historical financial behavior were derived so that the model can perform well..

3.2 Model Development

The following models were developed:

- **Logistic Regression**: Used as the baseline model.
- **Random Forest**: It is an ensemble learning method that is considered to be very robust and accurate.
- **Gradient Boosting Machine (GBM)**: An advanced ensemble method that builds models in a stage-wise fashion.
- **Ensemble Model**: This approach combines the predictions of logistic regression, random forest, and GBM via a soft voting mechanism.

3.3 Model Evaluation

Models were evaluated using the following metrics:

- **Accuracy**: It is the percentage of correctly predicted instances.
- **AUC-ROC**: It measures the ability of the model to distinguish between classes.
- **Precision, Recall, and F1-Score**: Used to evaluate the model's performance on both the majority (non-default) and minority (default) classes.

References

- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589-609. https://doi.org/10.2307/2978933

- Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

- Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124-136. https://doi.org/10.1016/j.ejor.2015.05.030

- Thomas, L. C., Crook, J. N., & Edelman, D. B. (2017). *Credit scoring and its applications* (2nd ed.). SIAM.