

# Configuration Manual

MSc Research Project  
Data Analytics

Bolormaa Mendbayar  
Student ID: X23176725

School of Computing  
National College of Ireland

Supervisor: Naushad Alam

National College of Ireland  
Project Submission Sheet  
School of Computing



<b>Student Name:</b>	Bolormaa Mendbayar
<b>Student ID:</b>	X23176725
<b>Programme:</b>	Data Analytics
<b>Year:</b>	2024
<b>Module:</b>	MSc Research Project
<b>Supervisor:</b>	Naushad Alam
<b>Submission Due Date:</b>	12/08/2024
<b>Project Title:</b>	Configuration Manual
<b>Word Count:</b>	XXX
<b>Page Count:</b>	6

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.

<b>Signature:</b>	
<b>Date:</b>	10th August 2024

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:**

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

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

<b>Office Use Only</b>	
Signature:	
Date:	
Penalty Applied (if applicable):	

# Configuration Manual

Bolormaa Mendbayar  
X23176725

## 1 Introduction

This configuration manual specifies the guidelines for processing and predicting the loan amount distribution based on the HMDA 2022 data. This manual also gives a clear description of the required hardware and software programs for the successful implementation. It also provides detailed instructions to help the users in setting up the analysis and in performing the analysis on the mortgage data using complex statistical and machine learning methods.

Note: All the code necessary to run the steps outlined in this manual is contained in the Jupyter Notebook file x23176725.ipynb can be found in the zipped artefact attached.

## 2 System Configuration

This section describes the hardware and software requirements used in this project in detail.

### 2.1 Hardware Requirements

Table 1 shows the hardware requirements for the system used in this project.

Table 1: Hardware Requirements

Hardware	Configurations
System	DELL XPS 13 7390
Processor	Intel(R) Core(TM) i7-10710U
Installed RAM	8.00 GB
Operating System	Windows 10 (64 Bits)
Hard Disk	256GB
Graphics Card	Intel(R) UHD Graphics (4GB)

Device specifications	
Device name	DESKTOP-ERDNT00
Processor	Intel(R) Core(TM) i7-10710U CPU @ 1.10GHz 1.61 GHz
Installed RAM	8.00 GB (7.79 GB usable)
Device ID	89B64479-CE57-42C8-B7A2-17B9CC46BCCF
Product ID	00326-30000-00001-AA526
System type	64-bit operating system, x64-based processor
Pen and touch	No pen or touch input is available for this display

Figure 1: System Configuration

Figure 1 shows the operating system configurations.

## 2.2 Software Requirements

Table 2 illustrates the detailed software requirements for the whole project. Figure 3 displays an example of Python code in Visual Studio Code.

To set up Visual Studio Code (VSCode) on Windows, the initial step is to download and install Python from <https://www.python.org/downloads> and then download and install VSCode from <https://code.visualstudio.com>. For detailed steps on how to install VSCode for Mac, check this link: <https://code.visualstudio.com/docs/setup/mac>.

Table 2: Software Requirements

Software	Version
Python	3.12 (64 Bits)
Visual Studio Code	1.60.0 or later (64 Bits)
Jupyter Notebook Extension	Latest version



Figure 2: Visual Studio Code Collaboratory with Python

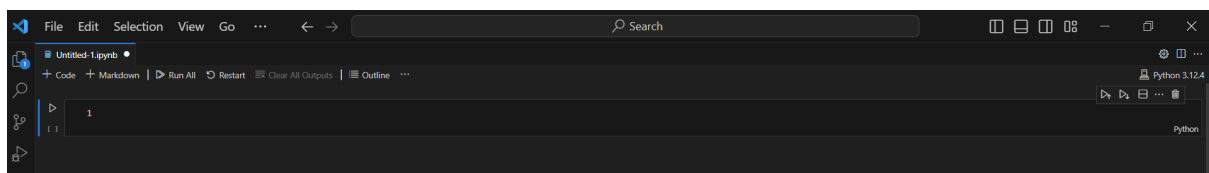


Figure 3: Python code files in Visual Studio Code

## 2.3 Libraries

- matplotlib
- pandas
- seaborn
- catboost
- lightgbm
- mpl\_toolkits
- plotly
- scipy
- sklearn
- statsmodels
- xgboost
- geopandas
- numpy

## 3 Project Implementation

### 3.1 Data Collection

Dataset link: <https://ffiec.cfpb.gov/data-publication/snapshot-national-loan-level-dataset/2022>

Cartographic Boundary Files: <https://www.census.gov/geographies/mapping-files/time-series/geo/kml-cartographic-boundary-files.html>

### 3.2 Methodology

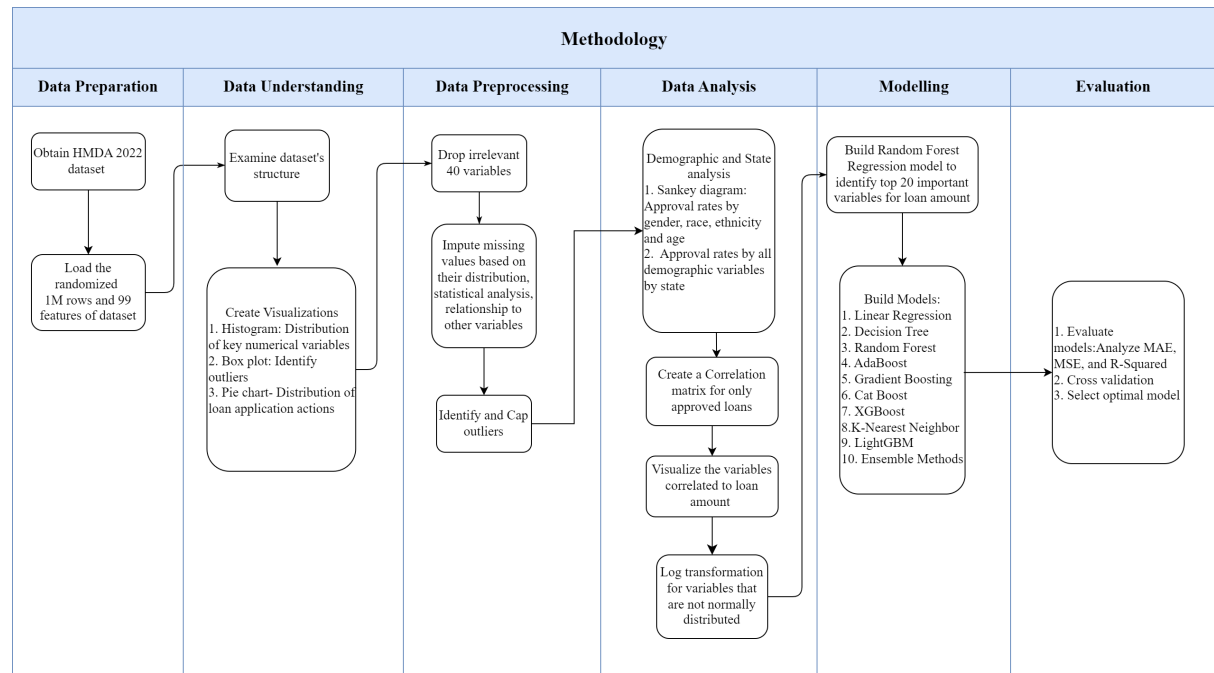


Figure 4: Research Methodology

## 4 Process Overview

### 4.1 Data Preparation

The dataset was shuffled using **pandas** and divided into chunks, each containing 1 million rows to facilitate more manageable processing. The code can be found in the x23176725.ipynb file, in the ‘1. Data Preparation’ section.

### 4.2 Data Understanding

This phase involved analyzing the dataset size, examining variable types, and identifying missing values. To understand the distribution of variables, the following visualizations were used:

- **Histograms:** Visualized the distribution of both numerical and categorical variables.
- **Pie Charts:** Displayed the proportion of various loan actions within the dataset.

This visualization code can be found in the x23176725.ipynb file, in the ‘2. Data Understanding’ section.

### 4.3 Data Preprocessing

Preprocessing involved imputing missing values and treating outliers:

- **Handling Missing Values:** The relevance of each feature to loan amount and approval was assessed, and irrelevant features were removed. Histograms were generated to analyze the distribution of variables with missing values, and correlation coefficients were calculated to identify relationships among variables. Based on these analyses, missing values were imputed.
- **Outlier Treatment:** Outliers were identified and capped based on the 5th and 95th percentiles of each numerical column.

This step’s code can be found in the x23176725.ipynb file, in the ‘3. Data Preprocessing’ section.

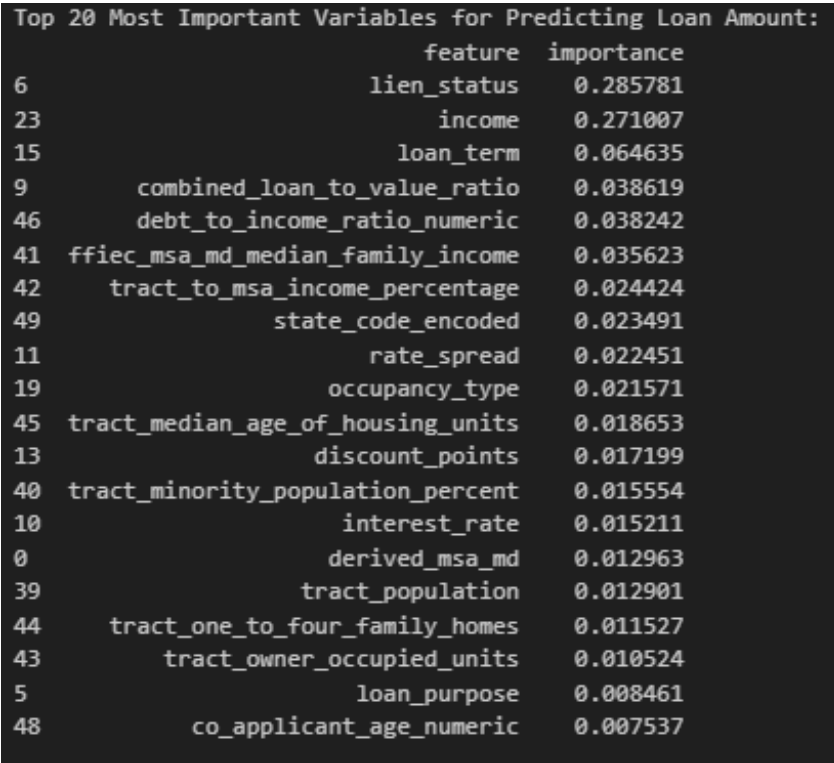
### 4.4 Data Analysis

- **Sankey Diagrams:** Used to analyze approval rates in demographic and financial biases at the national level.
- **State Analysis:** Conducted to examine approval rates in demographic and financial biases at the state level.
- **Correlation Matrix:** Generated to understand the relationships between loan amounts and other variables.

Visualizations and analysis code can be found in the x23176725.ipynb file, in the ‘4. Data Analysis’ section.

## 4.5 Modelling

**Feature Selection:** Random Forest was used for feature importance analysis. The top 20 most important variables for predicting loan amounts are shown in Figure 5.



	feature	importance
6	lien_status	0.285781
23	income	0.271007
15	loan_term	0.064635
9	combined_loan_to_value_ratio	0.038619
46	debt_to_income_ratio_numeric	0.038242
41	ffiec_msa_md_median_family_income	0.035623
42	tract_to_msa_income_percentage	0.024424
49	state_code_encoded	0.023491
11	rate_spread	0.022451
19	occupancy_type	0.021571
45	tract_median_age_of_housing_units	0.018653
13	discount_points	0.017199
40	tract_minority_population_percent	0.015554
10	interest_rate	0.015211
0	derived_msa_md	0.012963
39	tract_population	0.012901
44	tract_one_to_four_family_homes	0.011527
43	tract_owner_occupied_units	0.010524
5	loan_purpose	0.008461
48	co_applicant_age_numeric	0.007537

Figure 5: Top 20 most important variables for loan amount prediction

**Data Split:** The dataset was split into 80% for training and validation and 20% for testing. The 80% training data was further divided, with 25% reserved for validation. This resulted in a final split of 60% training, 20% validation, and 20% testing.

### Hyperparameters:

- **Decision Tree:** Random states of 30, 50, and 80 were used in Set 1, Set 2, and Set 3, respectively.
- **Random Forest, Gradient Boosting, AdaBoost, XGBoost, LightGBM:** Evaluated with 10, 20, and 40 estimators, using corresponding random states of 30, 50, and 80 across the three sets.
- **K-Nearest Neighbors:** Tested with 10, 20, and 40 neighbors in Set 1, Set 2, and Set 3, respectively.
- **CatBoost:** Evaluated with 10, 20, and 40 iterations, using random seeds of 30, 50, and 80.
- **Linear Regression:** Used as a baseline model without hyperparameter adjustments.

**Models:** The following 10 models were employed for loan amount prediction: Linear Regression, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, K-Nearest Neighbors, XGBoost, LightGBM, CatBoost, and an Ensemble method.

Feature Importance, and Modelling code can be found in the x23176725.ipynb file, in the ‘5. Modelling’ section.

## 4.6 Evaluation

Model performance was evaluated using the following metrics:

- **Mean Absolute Error (MAE)**
- **Mean Squared Error (MSE)**
- **R-squared**

For the ensemble method, the top three models such as CatBoost, XGBoost, and Random Forest were selected based on the highest R-squared and lowest error metrics. These models were configured with consistent parameters for a comparative analysis across all metrics.

Evaluation code can be found in the x23176725.ipynb file, in the same section as modelling.