National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Parth Patel
Student ID: X23166029

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Parth Patel
Student ID:	X23166029
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Noel Cosgrave
Submission Due Date:	12/12/2024
Project Title:	Configuration Manual
Word Count:	490
Page Count:	5

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:	
Date:	25th January 2025

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

Parth Patel
x23166029

1 Introduction

The main purpose of this research is to come up with an AI-based personalized loan pricing model for peer-to-peer lending platforms. Through the use of machine learning models, the project assesses the borrower's likelihood of defaulting and sets the interest rates that are risk-adjusted. All types of data pre-processing, model training, and evaluation were performed using Python in the Jupyter Notebook environment Ala'raj and Abbod (2016) Zhu and Yuan (2018).

This configuration manual gives you an idea of the hardware, environment, data sources, Python libraries, and models which have been used in this research to make it reproducible Feng and Dong (2018) Mai and Ma (2019).

2 Hardware

The research was conducted on the following hardware setup:

- **Device:** MacBook Air M1 (Apple Silicon)
- **Operating System:** macOS Monterey (Version 12.x)
- **Processor:** Apple M1 (8-core CPU)
- **RAM:** 8 GB
- **Storage:** 256 GB SSD

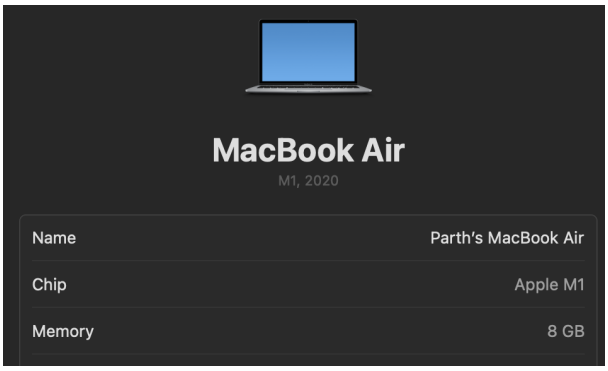

Figure 1: System Configurations

This configuration ensured sufficient computational power to handle the data set and execute machine learning models.

3 Environment

The project was developed in the Jupyter Notebook, a web-based computer program for the Python programming language. The Jupyter Notebook was installed through **pip**, Python's command line software installer - the external distribution software stack was not needed.

3.1 Environment Setup

1. **Install Python**: Python 3.9 was used for this project. Install Python from the official website: https://www.python.org/
2. **Install Jupyter Notebook**: Install Jupyter Notebook using the following command: `pip install notebook`
3. **Launch Jupyter Notebook**: Run the following command to start the environment: `jupyter notebook`

4 Data

LendingClub Corporation, one of the pioneering P2P (peer-to-peer) lending corporations, is the creator of the data set. The data is made up of loan records from 2007 to 2018 that include detailed information about the borrower and the loan.

4.1 Dataset Details

- **File:** `accepted_2007_to_2018Q4.csv`
- **Size:** Approximately 1.8 GB
- **Attributes:** 151 columns describing loan details, borrower profiles, and payment statuses.
- **Target Variable:** `loan_status` (Encoded as `Fully Paid` and `Charged Off`)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2260701 entries, 0 to 2260700
Data columns (total 31 columns):
 #   Column                Dtype
---  ------                -----
 0   loan_amnt             float64
 1   term                  object
 2   int_rate              float64
 3   installment           float64
 4   grade                 object
 5   sub_grade             object
 6   emp_title             object
 7   emp_length            object
 8   home_ownership        object
 9   annual_inc            float64
 10  verification_status   object
 11  issue_d               object
 12  loan_status           object
 13  purpose               object
 14  title                 object
 15  addr_state            object
 16  dti                   float64
 17  earliest_cr_line      object
 18  fico_range_low        float64
 19  fico_range_high       float64
 20  inq_last_6mths        float64
 21  open_acc              float64
 22  pub_rec               float64
 23  revol_bal             float64
 24  revol_util            float64
 25  total_acc             float64
 26  initial_list_status   object
 27  last_pymnt_amnt       float64
 28  application_type      object
 29  mort_acc              float64
 30  pub_rec_bankruptcies  float64
dtypes: float64(16), object(15)
memory usage: 534.7+ MB
```

Figure 2: Data Explanation

The data set was downloaded from the LendingClub official website:
https://www.lendingclub.com/info/download-data.action

5 Python Libraries

The following Python libraries were used for data processing, visualization, model development, and evaluation:

- **Data Manipulation:** `pandas`, `numpy`
- **Visualization:** `matplotlib`, `seaborn`
- **Machine Learning:** `scikit-learn`, `xgboost`, `lightgbm`, `tensorflow`
- **Oversampling:** `imbalanced-learn`
- **Model Interpretation:** `shap`, `lime`

5.1 Installation

Install the libraries using the following commands: **pip install pandas numpy matplotlib seaborn scikit-learn xgboost lightgbm tensorflow shap lime imbalanced-learn**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('ggplot')
import zipfile

# Pandas display options
pd.set_option('display.max_columns', 500)
pd.set_option('display.max_colwidth', 500)
pd.set_option('display.max_info_columns', 500)
pd.set_option('display.max_info_rows', 2000)
pd.set_option('display.expand_frame_repr', True)
pd.set_option('display.width', 2000)
pd.set_option('display.max_rows', 500)

import warnings
warnings.filterwarnings('ignore')
```

Figure 3: Import of libraries

6 Models

The project implemented the following machine learning models to evaluate borrower default probabilities and optimize loan pricing.

1. **Random Forest**
2. **XGBoost**
3. **LightGBM**
4. **Neural Network**

6.1 Model Building Process

1. **Data Splitting:** The data set was divided into training (70%) and testing (30%) subsets.
2. **Feature Selection:** Important features were selected using the Random Forest feature importance plot.
3. **Scaling:** StandardScaler was applied to normalize the feature ranges.
4. **Hyperparameter Tuning:** GridSearchCV was used for optimal parameter selection.

Random Forest Classifier

```
from sklearn.ensemble import RandomForestClassifier

rf_clf = RandomForestClassifier().fit(X_train, y_train)
rf_pred = rf_clf.predict(X_test)

from sklearn.metrics import classification_report, confusion_matrix

print(confusion_matrix(y_test, rf_pred))
print(classification_report(y_test, rf_pred))
```

```
[[9219  232]
 [ 171 1596]]
              precision    recall  f1-score   support

           0       0.98      0.98      0.98      9451
           1       0.87      0.90      0.89      1767

    accuracy                           0.96     11218
   macro avg       0.93      0.94      0.93     11218
weighted avg       0.96      0.96      0.96     11218
```

Figure 4: Random Forest Classifier Snippet

References

Ala'raj, M. and Abbod, M. F. (2016). A new hybrid ensemble credit scoring model based on classifiers consensus system approach, *Expert Systems with Applications* **64**: 36–55.

Feng, X., X. Z. Z. B. Q. J. and Dong, Y. (2018). Dynamic ensemble classification for credit scoring using soft probability, *Applied Soft Computing* **65**: 139–151.

Mai, F., T. S. L. C. and Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures, *European Journal of Operational Research* **274**(2): 743–758.

Zhu, B., Y. W. W. H. and Yuan, Y. (2018). A hybrid deep learning model for consumer credit scoring, *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, IEEE, pp. 205–208.