A Course Based Project Report on

RISK ANALYSIS FOR HOME CREDIT DEFAULT

Submitted to the

Department of CSE-(CyS, DS) and AI&DS

in partial fulfilment of the requirements for the completion of course

PYTHON PROGRAMMING LABORATORY (22ES2DS101)

BACHELOR OF TECHNOLOGY

IN

CSE- Data Science

Submitted by

B. ANJALI	23071A6706
L. YAMUNA	23071A6730
Y. BHAVANA SREE	23071A6765

Under the guidance of

Mr. G. Sathar

Assistant Professor

TAMASOMA JYOTIRGAMAYA
Estd.1995

Department of CSE-(CyS, DS) and AI&DS

VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI INSTITUTE OF ENGINEERING & TECHNOLOGY

An Autonomous Institute, NAAC Accredited with ‘A++’ Grade, NBA
Vignana Jyothi Nagar, Pragathi Nagar, Nizampet (S.O), Hyderabad – 500 090, TS, India

NOVEMBER-2024

VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

An Autonomous Institute, NAAC Accredited with 'A++' Grade, NBA Accredited for CE, EEE, ME, ECE, CSE, EIE, IT B. Tech Courses, Approved by AICTE, New Delhi, Affiliated to JNTUH, Recognized as "College with Potential for Excellence" by UGC, ISO 9001:2015 Certified, QS I GUAGE Diamond Rated
Vignana Jyothi Nagar, Pragathi Nagar, Nizampet(SO), Hyderabad-500090, TS, India

Department of CSE-(CyS, DS) and AI&DS

TAMASOMA JYOTIRGAMAYA
Estd.1995

CERTIFICATE

This is to certify that the project report entitled "**Risk Analysis For Home Credit Default**" is a bonafide work done under our supervision and is being submitted by **Miss. B. Anjali (23071A6706), Miss. L. Yamuna(23071A6730), Miss. Y. Bhavana sree (23071A6765)** in partial fulfilment for the award of the degree of **Bachelor of Technology** in **CSE-Data Science**, of the VNRVJIET, Hyderabad during the academic year 2024-2025.

Mr. G. Sathar	**Dr. T. SUNIL KUMAR**
Assistant Professor	Professor & HOD
Dept of CSE-(CyS, DS) and AI&DS	Dept of CSE-(CyS, DS)and AI&DS

Course based Projects Reviewer

VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

An Autonomous Institute, NAAC Accredited with ‘A++’ Grade,
Vignana Jyothi Nagar, Pragathi Nagar, Nizampet(SO), Hyderabad-500090, TS, India

Department of CSE-(CyS, DS) and AI&DS

TAMASOMA JYOTIRGAMAYA
Estd. 1995

DECLARATION

We declare that the course based project work entitled “**RISK ANALYSIS FOR HOME CREDIT DEFAULT**” submitted in the Department of **CSE-(CyS, DS) and AI&DS**, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, in partial fulfilment of the requirement for the award of the degree of **Bachelor of Technology in CSE-Data Science** is a bonafide record of our own work carried out under the supervision of **Mr. G. Sathar, Assistant Professor, Department of CSE-(CyS, DS) and AI&DS , VNRVJIET.** Also, we declare that the matter embodied in this thesis has not been submitted by us in full or in any part thereof for the award of any degree/diploma of any other institution or university previously.

Place: Hyderabad.

B. Anjali	**L. Yamuna**	**Y. Bhavana**
(23071A6706)	(23071A6730)	(23071A6765)

ACKNOWLEDGEMENT

We express our deep sense of gratitude to our beloved President, **Sri. D. Suresh Babu,** VNR Vignana Jyothi Institute of Engineering & Technology for the valuable guidance and for permitting us to carry out this project.

With immense pleasure, we record our deep sense of gratitude to our beloved Principal, **Dr. C.D Naidu,** for permitting us to carry out this project.

We express our deep sense of gratitude to our beloved Professor **Dr. T. Sunil Kumar**, Professor and Head, Department of CSE-(CyS, DS) and AI&DS , VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad-500090 for the valuable guidance and suggestions, keen interest and through encouragement extended throughout the period of project work.

We take immense pleasure to express our deep sense of gratitude to our belove Guide, **Dr. G. Sathar**, Assistant Professor in CSE-(CyS, DS) and AI&DS, VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad, for his/her valuable suggestions and rare insights, for constant source of encouragement and inspiration throughout my project work.

We express our thanks to all those who contributed for the successful completion of our project work.

B. ANJALI	23071A6706
L. YAMUNA	23071A6730
Y. BHAVANA SREE	23071A6765

TABLE OF CONTENTS

ABSTRACT

Financial institutions face persistent challenges in identifying and managing the risk of loan defaults, a problem that can significantly impact their profitability and operational sustainability. Loan defaults not only result in financial losses but also strain the credit ecosystem, creating ripple effects across lenders, borrowers, and the broader economy. Assessing an applicant's creditworthiness is a complex task that involves analyzing a wide range of factors, including demographic details, financial history, employment stability, and loan-specific attributes.

Traditional methods of risk assessment are often insufficient in addressing the complexities of modern credit evaluation due to the growing volume and diversity of data. Misjudging credit risk can result in either high-risk loans being approved or low-risk applicants being rejected, both of which can harm an institution's financial standing and reputation.

This project aims to address these issues by leveraging advanced data-driven techniques to analyze historical loan data. Specifically, the project focuses on:

1. Conducting a detailed exploratory data analysis (EDA) to uncover patterns and relationships between various factors influencing loan defaults.

2. Identifying the most significant predictors of default, including demographic, behavioral, and loan-related variables.

3. Building and fine-tuning predictive models using machine learning algorithms to estimate the likelihood of default for future applicants.

4. Providing actionable insights that financial institutions can use to optimize their loan approval processes, mitigate risks, and improve operational efficiency.

By combining data analysis and predictive modeling, this project seeks to develop a robust credit risk management framework that can help institutions reduce losses, enhance decision-making, and promote sustainable lending practices. It underscores the critical role of data science in modern financial risk management and highlights the potential of machine learning in transforming traditional credit evaluation systems.

CHAPTER-1

INTRODUCTION

The ability to accurately assess and manage credit risk is essential for financial institutions, particularly in the home credit sector, where defaults can have significant financial repercussions. Loan defaults not only result in monetary losses but also undermine the stability of the credit ecosystem, affecting both lenders and borrowers. In this context, data-driven solutions have emerged as powerful tools to address these challenges and enhance the efficiency of credit risk management systems.

This project focuses on analyzing historical loan data to gain insights into the factors contributing to loan defaults and developing predictive models to estimate the likelihood of default. By leveraging Python and its extensive suite of libraries for data processing, visualization, and machine learning, the project aims to build a robust framework for understanding credit risk. Exploratory Data Analysis (EDA) will help uncover patterns and relationships within the data, while machine learning algorithms will be employed to develop predictive models with high accuracy.

The insights gained from this analysis can empower financial institutions to make informed lending decisions, mitigate risks, and improve the overall efficiency of the loan approval process. Moreover, the use of data science techniques ensures a more objective and reliable assessment of creditworthiness, moving beyond traditional risk evaluation methods that may rely on subjective judgment or limited data.

The scope of this project is to build a data-driven solution for predicting the likelihood of loan defaults in the home credit sector. The project will leverage historical loan data, employing techniques such as Exploratory Data Analysis (EDA), machine learning modeling, and feature engineering to identify the key factors that contribute to defaults and to create predictive models. The goal is to develop a system that assists financial institutions in assessing the risk associated with each loan application, enabling them to make more informed, data-backed decisions. Historical loan data will be collected, which includes applicant information, financial history, and specific details about each loan.

CHAPTER-2

METHOD

Development Process

The first step in the project is **data collection and preprocessing**. Historical loan data will be collected, which includes applicant information, financial history, and specific details about each loan, such as loan amount, term, and repayment status. This data will likely contain missing values, inconsistencies, and categorical variables that need to be cleaned and encoded. Preprocessing will involve handling these issues by filling in missing values, encoding categorical features, and scaling or normalizing numerical features to prepare the data for further analysis.

Following this**, Exploratory Data Analysis (EDA)** will be performed. EDA is a critical step that will help us understand the data's underlying patterns and trends. During this phase, relationships between different features, such as applicant income, age, credit score, and loan amount, will be explored to uncover any significant correlations with loan defaults. Visualizations, such as histograms, boxplots, and scatter plots, will be used to illustrate these findings and present them in a way that is both insightful and easy to understand.

Once the data has been analyzed**, feature engineering** will be carried out. This involves creating new features from existing data, which could enhance the model's ability to make accurate predictions. Feature selection will also be applied to identify which features are most relevant to predicting loan defaults, ensuring the models focus on the most impactful variables.

The next phase of the project involves **model building and evaluation**. Various machine learning algorithms, including Logistic Regression, Decision Trees, Random Forests, and XGBoost, will be tested to build predictive models. These models will be evaluated based on several performance metrics, such as accuracy, precision, recall, F1-score, and AUC-ROC. The goal is to determine which model performs best in predicting the likelihood of default and how well it can generalize to new, unseen data.

In addition to predicting defaults, the project will aim to provide **risk assessment and decision support**. The models will output the probability of default for each loan applicant, which will help financial institutions assess the risk of approving each loan. This risk assessment can be used to inform decision-making, guiding lenders in approving or denying loans based on predicted risk levels.

Moreover, the insights gained from EDA will help improve the understanding of default risk factors, further enhancing decision-making strategies.

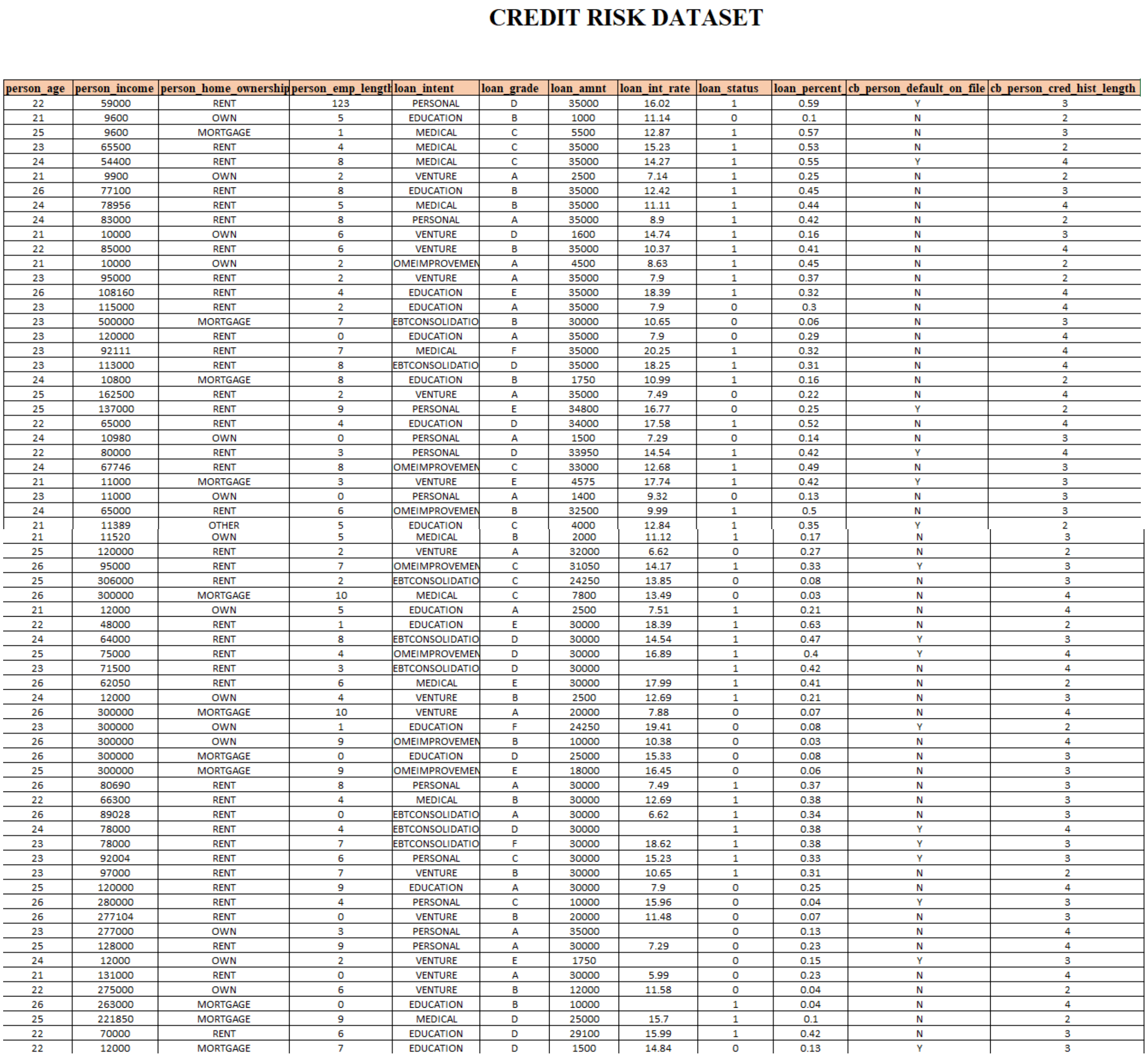
Tools Used

This project leverages a range of software tools and interfaces for effective data analysis and machine learning. **Python** serves as the primary programming language due to its versatility and extensive library support for tasks like data processing, exploratory analysis, and predictive modeling. **Jupyter Notebooks,** or optionally **Google Colab**, provide an interactive platform for writing and testing Python code, enabling real-time visualization of outputs like graphs and tables. Key libraries include **pandas for data manipulation, matplotlib and seaborn for data visualization**, and **scikit-learn** and **XGBoost** for implementing and optimizing machine learning models. Data storage is managed through **CSV files or SQL databases**, depending on dataset size and complexity.

On the hardware side, the project primarily requires a **computer or laptop** with at least 8GB of RAM and a capable processor, such as an Intel i5 or better. While a GPU is optional, it can expedite computations for large datasets or complex models. **Cloud computing resources** like Google Cloud or AWS may be employed for scalability if local hardware is insufficient. **External storage devices** or cloud storage solutions are useful for handling large data files or storing results, and a stable **internet connection** is essential for downloading libraries, accessing cloud services, and managing project data.

CREDIT RISK DATASET

Set Of Data:



CREDIT RISK DATASET

person_age	person_income	person_home_ownership	person_emp_lengtl	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_status	loan_percent	cb_person_default_on_file	cb_person_cred_hist_length
22	59000	RENT	123	PERSONAL	D	35000	16.02	1	0.59	Y	3
21	9600	OWN	5	EDUCATION	B	1000	11.14	0	0.1	N	2
25	9600	MORTGAGE	1	MEDICAL	C	5500	12.87	1	0.57	N	3
23	65500	RENT	4	MEDICAL	C	35000	15.23	1	0.53	N	2
24	54400	RENT	8	MEDICAL	C	35000	14.27	1	0.55	Y	4
21	9900	OWN	2	VENTURE	A	2500	7.14	1	0.25	N	2
26	77100	RENT	8	EDUCATION	B	35000	12.42	1	0.45	N	3
24	78956	RENT	5	MEDICAL	B	35000	11.11	1	0.44	N	4
24	83000	RENT	8	PERSONAL	A	35000	8.9	1	0.42	N	2
21	10000	OWN	6	VENTURE	D	1600	14.74	1	0.16	N	3
22	85000	RENT	6	VENTURE	B	35000	10.37	1	0.41	N	4
21	10000	OWN	2	OMEIMPROVEMEN	A	4500	8.63	1	0.45	N	2
23	95000	RENT	2	VENTURE	A	35000	7.9	1	0.37	N	2
26	108160	RENT	4	EDUCATION	E	35000	18.39	1	0.32	N	4
23	115000	RENT	2	EDUCATION	A	35000	7.9	0	0.3	N	4
23	500000	MORTGAGE	7	EBTCONSOLIDATIO	B	30000	10.65	0	0.06	N	3
23	120000	RENT	0	EDUCATION	A	35000	7.9	0	0.29	N	4
23	92111	RENT	7	MEDICAL	F	35000	20.25	1	0.32	N	4
23	113000	RENT	8	EBTCONSOLIDATIO	D	35000	18.25	1	0.31	N	4
24	10800	MORTGAGE	8	EDUCATION	B	1750	10.99	1	0.16	N	2
25	162500	RENT	2	VENTURE	A	35000	7.49	0	0.22	N	4
25	137000	RENT	9	PERSONAL	E	34800	16.77	0	0.25	Y	2
22	65000	RENT	4	EDUCATION	D	34000	17.58	1	0.52	N	4
24	10980	OWN	0	PERSONAL	A	1500	7.29	0	0.14	N	3
22	80000	RENT	3	PERSONAL	D	33950	14.54	1	0.42	Y	4
24	67746	RENT	8	OMEIMPROVEMEN	C	33000	12.68	1	0.49	N	3
21	11000	MORTGAGE	3	VENTURE	E	4575	17.74	1	0.42	Y	3
23	11000	OWN	0	PERSONAL	A	1400	9.32	0	0.13	N	3
24	65000	RENT	6	OMEIMPROVEMEN	B	32500	9.99	1	0.5	N	3
21	11389	OTHER	5	EDUCATION	C	4000	12.84	1	0.35	Y	2
21	11520	OWN	5	MEDICAL	B	2000	11.12	1	0.17	N	3
25	120000	RENT	2	VENTURE	A	32000	6.62	0	0.27	N	2
26	95000	RENT	7	OMEIMPROVEMEN	C	31050	14.17	1	0.33	Y	3
25	306000	RENT	2	EBTCONSOLIDATIO	C	24250	13.85	0	0.08	N	3
26	300000	MORTGAGE	10	MEDICAL	C	7800	13.49	0	0.03	N	4
21	12000	OWN	5	EDUCATION	A	2500	7.51	1	0.21	N	4
22	48000	RENT	1	EDUCATION	E	30000	18.39	1	0.63	N	2
24	64000	RENT	8	EBTCONSOLIDATIO	D	30000	14.54	1	0.47	Y	3
25	75000	RENT	4	OMEIMPROVEMEN	D	30000	16.89	1	0.4	Y	4
23	71500	RENT	3	EBTCONSOLIDATIO	D	30000		1	0.42	N	4
26	62050	RENT	6	MEDICAL	E	30000	17.99	1	0.41	N	2
24	12000	OWN	4	VENTURE	B	2500	12.69	1	0.21	N	3
26	300000	MORTGAGE	10	VENTURE	A	20000	7.88	0	0.07	N	4
23	300000	OWN	1	EDUCATION	F	24250	19.41	0	0.08	Y	2
26	300000	OWN	9	OMEIMPROVEMEN	B	10000	10.38	0	0.03	N	4
26	300000	MORTGAGE	0	EDUCATION	D	25000	15.33	0	0.08	N	3
25	300000	MORTGAGE	9	OMEIMPROVEMEN	E	18000	16.45	0	0.06	N	3
26	80690	RENT	8	PERSONAL	A	30000	7.49	1	0.37	N	3
22	66300	RENT	4	MEDICAL	B	30000	12.69	1	0.38	N	3
26	89028	RENT	0	EBTCONSOLIDATIO	A	30000	6.62	1	0.34	N	3
24	78000	RENT	4	EBTCONSOLIDATIO	D	30000		1	0.38	Y	4
23	78000	RENT	7	EBTCONSOLIDATIO	F	30000	18.62	1	0.38	Y	3
23	92004	RENT	6	PERSONAL	C	30000	15.23	1	0.33	Y	3
23	97000	RENT	7	VENTURE	B	30000	10.65	1	0.31	N	2
25	120000	RENT	9	EDUCATION	A	30000	7.9	0	0.25	N	4
26	280000	RENT	4	PERSONAL	C	10000	15.96	0	0.04	Y	3
26	277104	RENT	0	VENTURE	B	20000	11.48	0	0.07	N	3
23	277000	OWN	3	PERSONAL	A	35000		0	0.13	N	4
25	128000	RENT	9	PERSONAL	A	30000	7.29	0	0.23	N	4
24	12000	OWN	2	VENTURE	E	1750		0	0.15	Y	3
21	131000	RENT	0	VENTURE	A	30000	5.99	0	0.23	N	4
22	275000	OWN	6	VENTURE	B	12000	11.58	0	0.04	N	2
26	263000	MORTGAGE	0	EDUCATION	B	10000		1	0.04	N	4
25	221850	MORTGAGE	9	MEDICAL	D	25000	15.7	1	0.1	N	2
22	70000	RENT	6	EDUCATION	D	29100	15.99	1	0.42	N	3
22	12000	MORTGAGE	7	EDUCATION	D	1500	14.84	0	0.13	Y	3

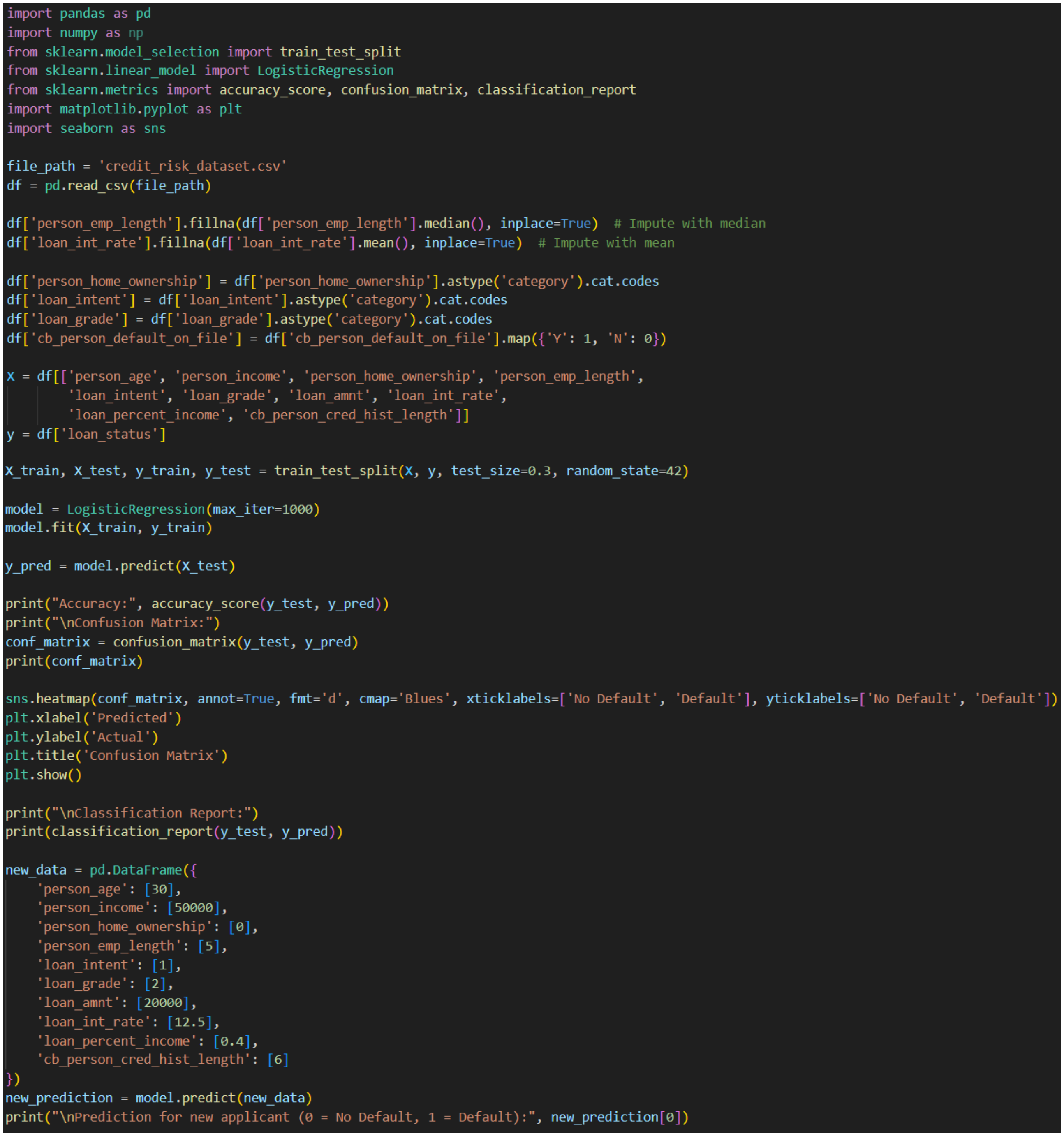
LINK TO DOWNLOAD DATASET:

https://www.kaggle.com/datasets/laotse/credit-risk-dataset

CHAPTER-3

CODE

PYTHON:

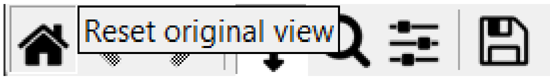


```python
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
import matplotlib.pyplot as plt
import seaborn as sns

file_path = 'credit_risk_dataset.csv'
df = pd.read_csv(file_path)

df['person_emp_length'].fillna(df['person_emp_length'].median(), inplace=True)  # Impute with median
df['loan_int_rate'].fillna(df['loan_int_rate'].mean(), inplace=True)  # Impute with mean

df['person_home_ownership'] = df['person_home_ownership'].astype('category').cat.codes
df['loan_intent'] = df['loan_intent'].astype('category').cat.codes
df['loan_grade'] = df['loan_grade'].astype('category').cat.codes
df['cb_person_default_on_file'] = df['cb_person_default_on_file'].map({'Y': 1, 'N': 0})

X = df[['person_age', 'person_income', 'person_home_ownership', 'person_emp_length',
        'loan_intent', 'loan_grade', 'loan_amnt', 'loan_int_rate',
        'loan_percent_income', 'cb_person_cred_hist_length']]
y = df['loan_status']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nConfusion Matrix:")
conf_matrix = confusion_matrix(y_test, y_pred)
print(conf_matrix)

sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues', xticklabels=['No Default', 'Default'], yticklabels=['No Default', 'Default'])
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix')
plt.show()

print("\nClassification Report:")
print(classification_report(y_test, y_pred))

new_data = pd.DataFrame({
    'person_age': [30],
    'person_income': [50000],
    'person_home_ownership': [0],
    'person_emp_length': [5],
    'loan_intent': [1],
    'loan_grade': [2],
    'loan_amnt': [20000],
    'loan_int_rate': [12.5],
    'loan_percent_income': [0.4],
    'cb_person_cred_hist_length': [6]
})
new_prediction = model.predict(new_data)
print("\nPrediction for new applicant (0 = No Default, 1 = Default):", new_prediction[0])
```

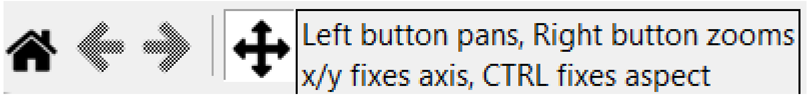
CHAPTER-4

TEST CASES/ OUTPUT

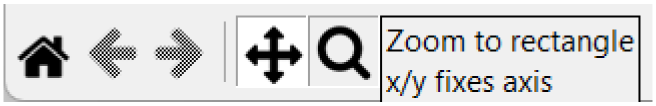
Basic Buttons:



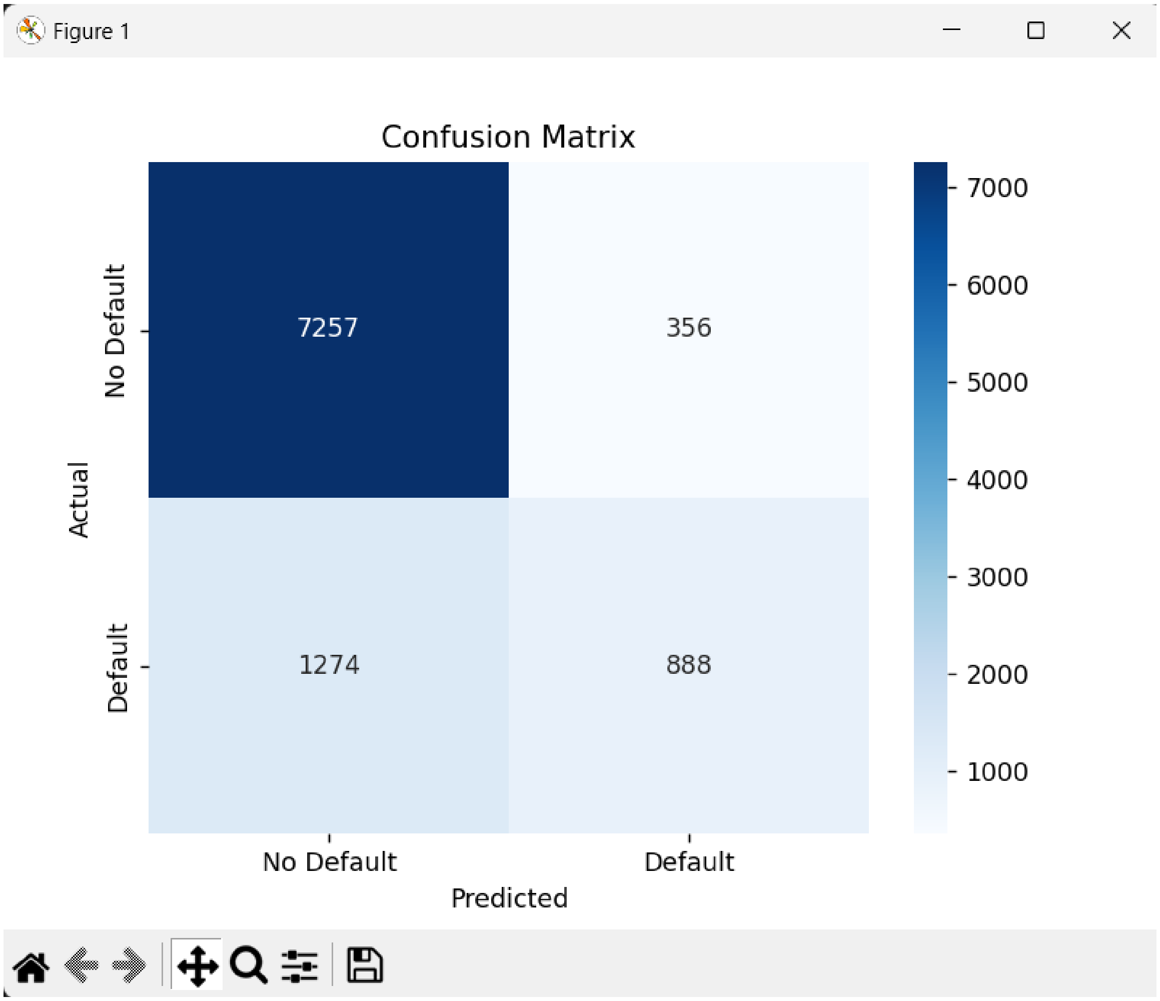
Reset original view



Left button pans, Right button zooms
x/y fixes axis, CTRL fixes aspect

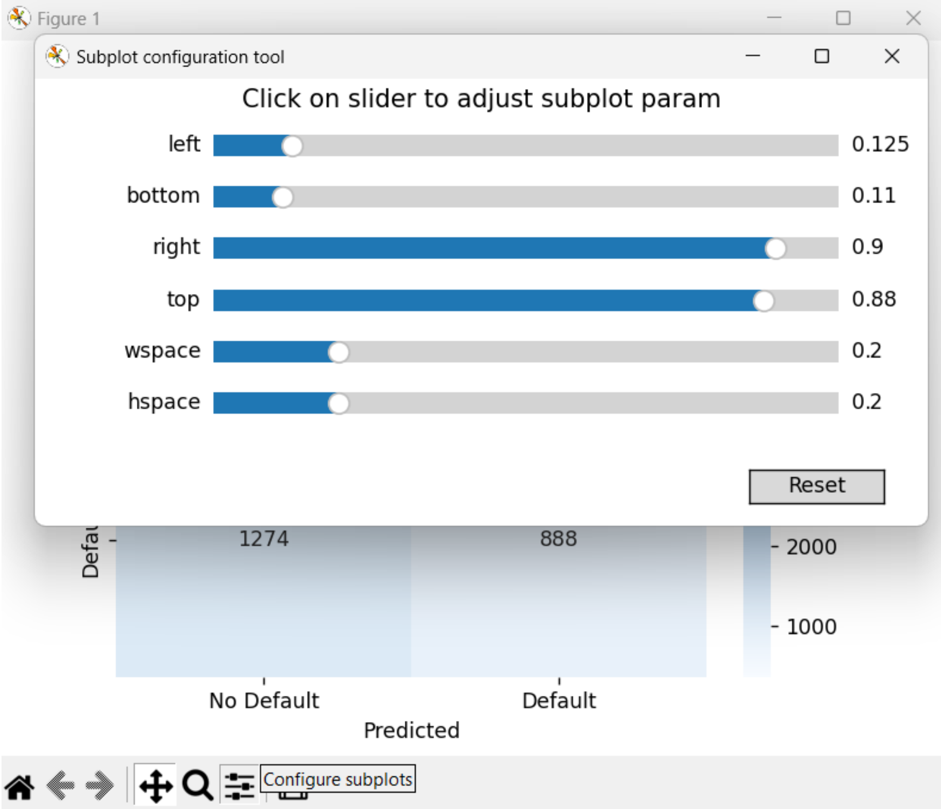


Zoom to rectangle
x/y fixes axis



Actual \ Predicted	No Default	Default
No Default	7257	356
Default	1274	888

Figure 1 - Positive scale matrix



Click on slider to adjust subplot param

left 0.125
bottom 0.11
right 0.9
top 0.88
wspace 0.2
hspace 0.2

Figure 2 - Sub plots

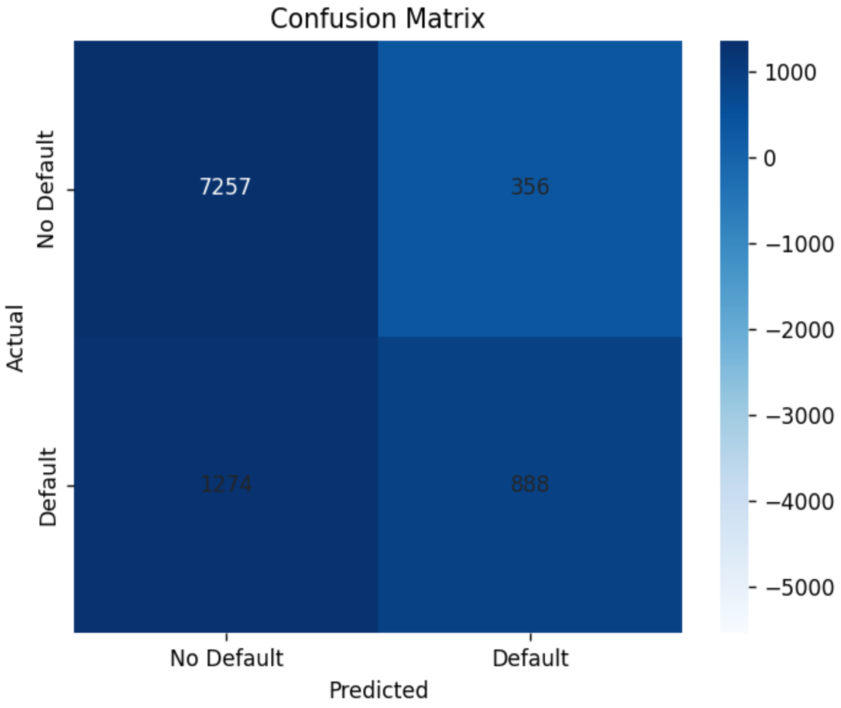
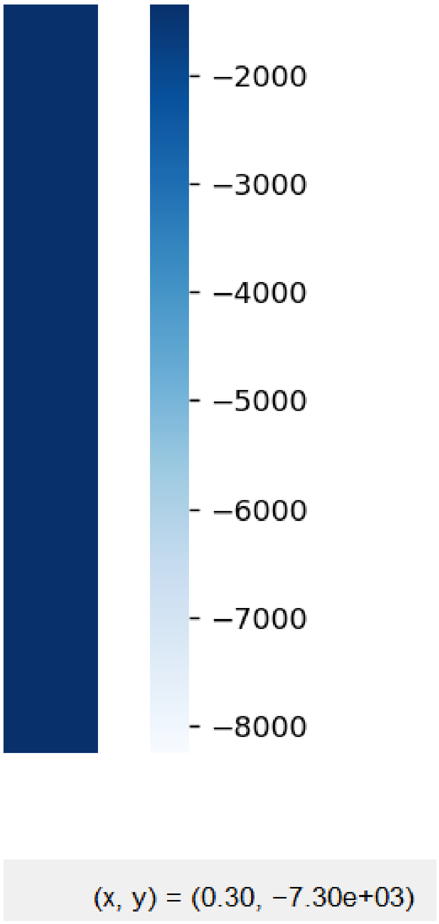
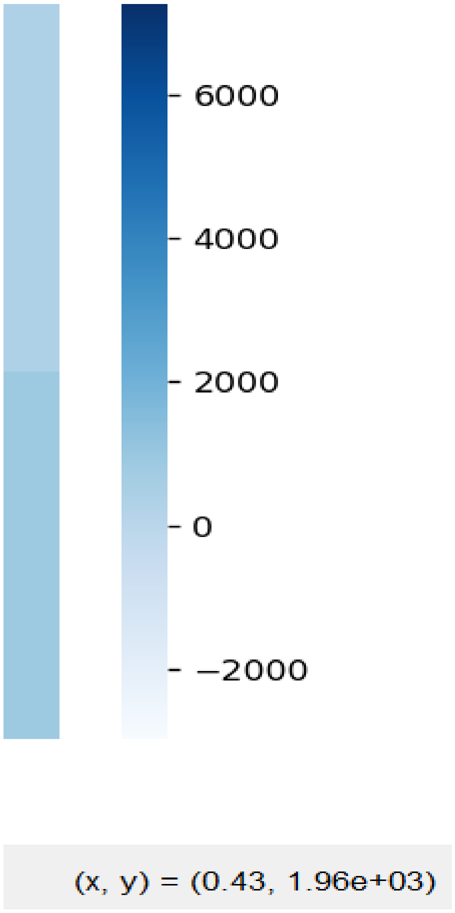


Figure 3 - Negative scale matrix


−2000
−3000
−4000
−5000
−6000
−7000
−8000
(x, y) = (0.30, −7.30e+03)


6000
4000
2000
0
−2000
(x, y) = (0.43, 1.96e+03)

CHAPTER-5

RESULTS

This project successfully addresses the critical challenge of assessing credit default risk for home loans by combining exploratory data analysis (EDA) and predictive modelling. The analysis provides valuable insights into the factors most strongly associated with loan default, such as applicant demographics, financial history, and loan characteristics. These insights can help financial institutions better understand the risks involved and make more informed lending decisions.

The predictive models developed in this project demonstrate their effectiveness in estimating the probability of default with a high degree of accuracy. By integrating these models into their decision-making processes, financial institutions can:

1. Reduce financial losses by identifying high-risk applicants.
2. Improve the efficiency of loan approval workflows.
3. Design strategies to mitigate credit risk and enhance profitability.

Overall, this project highlights the importance of leveraging data-driven techniques to improve credit risk management and support sustainable financial growth. Future work can involve refining the models using real-time data, exploring advanced machine learning algorithms, and incorporating external macroeconomic factors to enhance predictive accuracy.

CHAPTER – 6

SUMMARY

This project focuses on the critical task of analysing and predicting the risk of home credit loan defaults, an area that significantly impacts the financial sector. Financial institutions face considerable challenges in determining the likelihood that a borrower will default on a loan. Misjudging this risk can lead to financial losses, inefficiencies, and instability in the credit ecosystem. The objective of this project is to apply Exploratory Data Analysis (EDA) and predictive modelling techniques to a dataset of historical home loan applications, identifying key factors that influence loan defaults and developing machine learning models to predict the likelihood of default for future applicants.

The project begins with data collection and preprocessing, where historical loan data is cleaned and prepared for analysis. This includes handling missing values, encoding categorical variables, and normalizing or scaling numerical features. The next step involves performing Exploratory Data Analysis (EDA) to examine the relationships between applicant demographics, financial history, loan- specific attributes, and default outcomes. EDA helps to uncover patterns, correlations, and trends in the data, which are crucial for building predictive models. Various visualizations and statistical methods will be used to understand the underlying structure of the dataset, highlighting factors that have a significant impact on loan defaults.

Following the EDA, the project develops and evaluates predictive models using machine learning techniques such as Logistic Regression, Decision Trees, Random Forests, and XGBoost. These models are trained on the historical loan data and tested for accuracy and effectiveness in predicting default risk. By leveraging Python and key data science libraries, the project provides an end-to-end solution for risk assessment in home credit loans.

REFERENCES

1) *Home Credit Default Risk: Kaggle Dataset Overview.* Retrieved from https://www.kaggle.com
2) *Machine Learning Approaches to Credit Scoring.* Journal of Financial Risk Analytics, Vol. 4, Issue 2, 2022.
3) *Credit Risk Modeling Using Logistic Regression and Random Forest.* International Journal of Data Science Research, 2021.
4) *Feature Engineering Techniques for Financial Risk Assessment.* Proceedings of the 2023 IEEE Financial Analytics Symposium.
5) *Home Credit Risk: A Comparative Analysis of Models.* Data Science Journal, 2020.
6) *Interpreting Credit Default Models with SHAP.* Retrieved from https://towardsdatascience.com
7) *The Role of Socioeconomic Variables in Credit Scoring Models.* Global Financial Insights,