

---

# PROTOTYPE REPORT

---

## TABLE OF CONTENTS

CHAPTERS	PAGE NO.
<b>Title Page</b>	<b>01</b>
<b>Table of Contents</b>	<b>02</b>
<b>Abstract</b>	<b>04</b>
<b>1. Executive Summary</b>	<b>05</b>
1.1 Objective of the prototype	<b>05</b>
1.2 Expected Business Impact	<b>05</b>
<b>2. Introduction</b>	<b>08</b>
2.1 Background: Credit Risk in the Financial Industry	<b>08</b>
2.2 Importance of Alternative Data in Risk Assessment	<b>08</b>
<b>3. Business Understanding</b>	<b>09</b>
3.1 Use Case Overview: Probability of Default (PD) Estimation	<b>09</b>
3.2 Defaulter Search & Tracking Objectives	<b>10</b>
<b>4. Proposed Methodology</b>	<b>11</b>
4.1 Data Acquisition	<b>11</b>
4.2 Data Sources (Conventional & Alternative Data)	<b>12</b>
4.3 Data Description	<b>13</b>
<b>5. Exploratory Data Analysis (EDA)</b>	<b>20</b>
5.1 Descriptive Statistics	<b>20</b>

---

5.2 Missing Value Analysis	22
5.3 Correlation Analysis	23
5.4 Insights from EDA	24
<b>6. Feature Engineering</b>	<b>25</b>
6.1 Encoding Categorical Variables	25
6.2 Scaling and Normalization	25
6.3 Dimensionality Reduction	26
<b>7. Model Development</b>	<b>27</b>
7.1 Problem Formulation (Multi-output Regression for PD & Credit Score)	27
7.2 Baseline Model Selection (Linear Regression)	29
<b>8. Model Evaluation</b>	<b>35</b>
8.1 Evaluation Metrics (MAE, RMSE, $R^2$ )	35
8.2 Residual Analysis	37
8.3 Predicted vs Actual Plots	42
8.4 Cross-Validation Results	50
8.5 Model Comparison Table	51
<b>9. Conclusion</b>	<b>52</b>
<b>10. Future Scope</b>	<b>53</b>
<b>11. References</b>	<b>55</b>

---

## ABSTRACT

A major group of people and micro, small, and medium-sized enterprises (MSMEs) in developing markets are out of formal banking channels because they have sparse or no credit histories. This is because the conventional credit-scoring systems hinge on past financial history in terms of credit bureau reports and bank statements, which inherently penalize thin-file or first-time borrowers. The resulting gap not only limits the availability of credit to these groups but also denies growth potential to lenders. This prototype meets the challenge by creating a credit risk management model that combines alternative data sources such as digital transaction behaviour, utility bill payment history, mobile usage statistics, and other non-traditional factors with traditional financial characteristics. The model predicts the default probability (PD) and includes defaulter search and tracking functionality to provide early intervention in advance of delinquency. Focused on transparency, scalability, and compliance, the method seeks to enhance the accuracy of credit judgments, increase lending outreach to underpenetrated segments, and enhance portfolio quality while staying within data privacy standards. By using alternative data, the suggested solution facilitates inclusive finance, lowers non-performing assets, and provides banks and financial institutions with an operable route to extend credit responsibly to underbanked markets.

---

# EXECUTIVE SUMMARY

## 1.1 Objective of the Prototype

The goal of this prototype is the creation of a new next-generation credit risk management model. One that uses alternative data sources to ascertain a borrower's ability to default, with such a focus on underbanked and thin-file segments. A significant portion of the population in many emerging markets is excluded from the lending mainstream for not having an extended credit history. Hence, traditional models were to a great extent based on historical financial records, for example, credit bureau scores, bank statements, which in theory discriminate against someone, small businesses, etc., who do not have these records.

By using alternative indicators like mobile phone usage patterns, utility bills payment histories, or in instances behavioral or transactional data, it aims at entering creditworthiness in a more inclusive, fair, and accurate manner. Another feature of the prototype is that it will be able to search and track potential defaulters to intervene at the risk level early

## 1.2 Expected Business Impact

- **Expanded Lending Opportunities:** Enables banks to serve underbanked individuals and MSMEs by assessing creditworthiness through alternative data, creating a newer revenue stream for banks.
- **Improved Risk Prediction:** Improved accuracy in PD estimates means that fewer NPAs exist in the portfolios and an increase in portfolio stability.
- **Faster Credit Decisions:** Automated scoring and defaulter tracking minimize loan processing times and effort involved with manual checks.
- **Optimized Capital Allocation:** Better segmentation based on risks also allows banks to assign interest rates, maximum loan amounts, and loan terms based on the true risk profile of the borrower.
- **Regulatory Compliance & Transparency:** Model design forced greater explainability and fairness that respect data privacy law and lending regulations.

- 
- More access to credit for MSMEs: Alternate data sources recognize the credit potential of MSMEs sans extensive financial histories, enabling them to access formal financing.
  - Fairer loan terms: More accurate risk profiling allows well-deserving MSMEs to avail of lower interest rates and flexible repayment options.
  - Development of Business: Linking working capital in a timely manner to allowing MSMEs to grow, create jobs, and foster economic development.
  - Credit History Building: Enterprise and personal repayment predictably successful cases to activate improved credit profiles for further loan taking.
  - Symbiotic Ecosystem: Bankers are presented the opportunity of safe growth and portfolio strength while the MSMEs are given a chance to receive capital needed to scale up in a more sustainable manner.

---

# INTRODUCTION

In many emerging markets, access to formal credit remains a challenge for large segments of the population and small businesses. Many individuals and micro, small, and medium enterprises (MSMEs) are either underbanked or completely outside the formal banking system. Traditional credit assessment methods rely heavily on historical financial records such as bank statements, collateral history, and credit bureau scores. These methods do not adequately evaluate these groups. As a result, potentially creditworthy borrowers often get denied access to funding, and lenders miss chances to grow their customer base. This prototype aims to develop a credit risk management model that uses both conventional and alternative data sources to estimate the **probability of default (PD)** for borrowers. The model seeks not only to predict PD accurately but also to identify and monitor potential defaulters before risk exposure increases. By using machine learning techniques such as gradient boosting and neural networks, the solution is designed to deliver better accuracy, transparency, and scalability.

---

## **2.1 Background: Credit Risk in the Financial Industry**

Credit risk refers to the potential loss a lender may experience if a borrower does not meet their repayment obligations. In the financial industry, accurately assessing this risk is crucial for maintaining a healthy loan portfolio, meeting regulatory requirements, and ensuring banking stability. Traditional credit risk assessment heavily relies on historical data such as repayment records, credit bureau reports, bank statements, and collateral evaluations. While these methods work well for established borrowers, they often do not adequately assess individuals and small businesses with limited or no formal credit history.

This gap is especially clear in emerging markets, where a large percentage of the population remains underbanked or unbanked. For financial institutions, this means missed chances to expand their lending portfolios. For individuals and MSMEs, it leads to limited access to the funds necessary for growth and economic participation.

## **2.2 Importance of Alternative Data in Risk Assessment**

In recent years, the rise of digital services and mobile technology has provided new data sources that can offer valuable insights into borrower behavior. Alternative data includes non-traditional information such as mobile phone usage, utility bill payments, e-commerce transactions, rental payment history, social media activity, and digital wallet usage patterns.

These data points can provide a broader view of a borrower's financial reliability, particularly when traditional records are lacking. For instance, consistent utility payments or stable mobile recharge habits may signal a borrower's ability and willingness to meet obligations, even without a formal credit history. Using such data not only improves predictive accuracy but also supports financial inclusion by enabling fairer assessments for those previously left out of the lending system.



---

# BUSINESS UNDERSTANDING

## 3.1 Use Case Overview: Probability of Default (PD) Estimation

The Probability of Default (PD) is an important metric in credit risk management. It shows the chance that a borrower will fail to repay their loans on time, usually within a year. In banking, PD is crucial for loan pricing, calculating capital reserves, determining provisioning needs, and managing overall portfolio risk. Traditionally, PD is mainly based on historical repayment patterns, credit bureau scores, and income verification. This method works well for established borrowers but does not hold up for thin-file customers, like individuals or small and medium-sized enterprises (MSMEs) with limited formal credit histories. This issue creates a gap in risk assessment and leaves out many creditworthy applicants from the lending process.

This project aims to solve that problem by incorporating alternative data into the PD estimation framework. Examples of this data include:

- Behavioral patterns, such as utility bill payments, prepaid mobile recharge patterns, and e-commerce purchase history.
- Transactional activity, including mobile wallet usage, frequency of digital payments, and transaction volumes.
- Operational indicators for MSMEs, such as supplier payment timeliness, inventory turnover rates, and customer payment cycles.

By combining these signals with traditional credit data, the model intends to provide more accurate and inclusive PD predictions. This approach allows banks to offer credit responsibly to borrowers who were previously ignored.

---

## 3.2 Defaulter Search & Tracking Objectives

While a reliable PD estimation is essential for approving loans, ongoing risk monitoring is also critical for keeping a healthy loan portfolio. The Defaulter Search & Tracking feature in this project acts as a real-time early warning system.

Key objectives include:

### 1. Risk Flagging at Origination

- Using PD thresholds (for example,  $PD > 0.6$ ) to categorize borrowers into risk tiers at the application stage.
- Ensuring that high-risk applicants get extra scrutiny or adjusted loan terms.

### 2. Continuous Behavioral Monitoring

- Watching borrower activity after loan disbursement, such as changes in transaction frequency, missed utility payments, or sudden drops in digital wallet balances.
- Identifying early signs of financial distress that could lead to default.

### 3. Searchable Risk Registry

- Keeping a centralized database of flagged accounts with detailed risk scores, historical alerts, and intervention records.
- Allowing access across departments for collections, customer relationship teams, and compliance officers.

### 4. Proactive Intervention

- Initiating targeted actions like payment reminders, restructuring offers, or credit limit reductions before a payment is missed.
- Prioritizing recovery strategies for accounts most at risk of default.

This feature shifts defaulter management from being reactive (after missed payments) to proactive (before default happens). This change helps reduce Non-Performing Assets (NPAs) and improves recovery rates.

---

# Proposed Methodology

We have followed the following order to develop this prototype model

## 4.1 Data Acquisition

The data acquisition process for alternative credit scoring draws upon multiple sources of data to include conventional and alternative measures of borrower creditworthiness. Conventional data has been sourced from microfinance institutions, fintech companies, and consumer credit repositories; loans which show repayment history, demographic data, and structured financial transactions (Kumar & Babu, 2025). In addition, there are alternative datasets to consider, which include psychometric assessments, patterns in email usage, and digital traces that are used to profile borrowers with little to no formal credit history..

In addition, social network information has emerged as a critical component for enhancing predictive performance. **Niu et al. (2019)** demonstrate that mobile phone-derived social network quality, stability, and exposure provide statistically significant insights into default risk, especially in peer-to-peer lending contexts. Similarly, **Alamsyah et al. (2025)** highlight the use of LinkedIn-derived demographics, personality traits, psycholinguistics, and professional network data to assess creditworthiness for individuals lacking conventional financial records or collateral.

In their analysis of the Indian financial ecosystem, **Bokade et al. (2025)** stress the benefits that behavioral data could provide (e.g., mobile recharge frequency, utility bill pay punctuality, Unified Payments Interface (UPI) transaction trends as substitutes for financial discipline and digital engagement. These data would provide insight, especially for consumers engaging with credit products for the first time in rural or underdeveloped spaces.

In addition, this holistic approach creates borrower profiles through the acquisition of datasets with different sources and formats while ensuring the integrity of the data through common step of preprocessing it, which includes identifying outliers, normalizing, and imputing data with missing values.

---

## 4.2 Data Sources (Conventional & Alternative Data)

The creation of inclusive and fair credit scoring models depends on the inclusion of a variety of data sources, including both conventional and alternative data sets.

### **Conventional Data Sources:**

Legacy data basically comprise structured economic and demographic data collected from established institutions like microfinance institutions, fintech sites, consumer credit databases, and formal banking systems. Such datasets commonly involve loan payment histories, credit bureau data, income information, debt-to-income values, and borrower profiles (Kumar & Babu, 2025). Such data continue to be vital in generating a working knowledge of borrower behavior, particularly for those who have built up their credit profiles.

### **Alternative Data Sources:**

As the limitations of conventional data in evaluating "new-to-credit" or underbanked consumers have become apparent, alternative sources of data have emerged into focus. Some of these alternative data sources are psychometric testing, online traces, email patterns, and non-fintech transactions like rent and utility bill payments (Kumar & Babu, 2025). Social network data constitute a rich alternative dataset. **Niu et al. (2019)** illustrate how mobile phone based metrics social network quality, stability, and exposure can improve default prediction accuracy in peer-to-peer lending platforms significantly. **Alamsyah et al. (2025)** extend this strategy further by utilizing LinkedIn-based demographics, personality, psycholinguistics, and professional network connections to evaluate the creditworthiness of unsecured or traditional credit-record-less individuals.

In the Indian context, **Bokade et al. (2025)** emphasize the applicability of behavioral data like frequency of mobile recharges, utility bill payment punctuality, and patterns of Unified Payments Interface (UPI) transactions as surrogates of financial responsibility and digital activity. These metrics prove particularly useful for the rural population and gig economy laborers, where informal or cash-based economic activity is prevalent.

---

By integrating historical accounting statements with non-traditional datasets, credit scoring algorithms can more accurately capture a well-rounded, context-driven representation of borrowers, which in turn enhances predictability and extends credit access to varying socioeconomic ecosystems.

### 4.3 Data Description

The Dataset contains the following features which contains both traditional and alternative features which will be helpful for determining of PoD (Probability of Default) and possible credit score .

1. Age

Age can provide clues to financial maturity and borrowing behavior. Older people have, probably, longer credit histories. If they have not had delinquencies in their account, lenders have an extensive body of work to examine when gauging reliability. Younger applicants have shorter credit histories, which makes gauging reliability less certain.

2. Gender

Gender itself is not a determinant of creditworthiness but may correlate with statistical repayment trends. Responsible lending makes certain that its used as simply a demographic factor that might relate to use, not as a bias. Use may sometimes, with other factors combined, reveal behavior trends in spending or saving.

3. Marital\_Status

married applicants may be able to show greater financial stability because a married applicant has a combined income, or shared costs which may increase the ability to repay, but also has costs that must also be considered. Single applicants likely have a lack of dependants and any disposable income would be recognized in their income household.

4. Education\_Level

Higher education often means better job prospects and a higher rate of income. Education may also relate to financial literacy, including budgeting and management of debt.

5. Employment\_Status

Stable, full-time employment decreases credit risk as it produces predictable income. Self-employed or contract- based earners may be subject to inconsistent variability of cash flow which makes repayment less predictable. Lenders weigh the stability and reliability of the type of employment heavily.

---

## 6. Occupation

Occupations exhibit a diversity of distinct income patterns, job security levels, and career opportunities. For example, many government jobs imply some level of stability, while commission-based work comes with a variety of risk. This is useful in predicting future stability of income.

## 7. Annual\_Income

In general, a larger yearly income lowers the debt-to-income ratio and improves repayment capabilities. It can also act as a buffer against unforeseen costs. However, without taking spending patterns into account, income is insufficient on its own.

## 8. Monthly\_Income

For repayment plans to be predictable, monthly income must be steady and regular. Variability in monthly income may be a warning sign for possible late payments. This measure aids in determining if EMI-based loans are affordable.

## 9. Household\_Size

Higher living expenses are frequently associated with larger households, which might lower the percentage of income available for debt repayment. Financial flexibility may be greater for smaller households. When evaluating budget pressure, lenders take it into account in addition to income.

## 10. Dependents

Dependents raise continuous costs, particularly if they are children or elderly family members. Repayment capacity may be strained by having more dependents, particularly if income is fixed. When assessing discretionary income, this component is crucial.

## 11. Home\_Ownership

A residence can be used as collateral if necessary because it shows asset ownership and financial stability. Renters could have less material possessions and greater monthly responsibilities.

## 12. Residence\_Type

Staying in one place for an extended period of time suggests stability and a decreased danger of moving. Frequent movements can generate fears about volatility in employment or finances.

## 13. Credit\_History\_Length

Repayment behavior can be predicted more accurately with a longer credit history. Lenders are forced to depend more on recent activities because to the uncertainty created by short histories.

---

#### 14. Number\_of\_Credit\_Cards

While having multiple credit cards can demonstrate financial flexibility, having too many could indicate overstretching. Available credit data is limited by few or no cards. The ideal usage is balanced.

#### 15. Number\_of\_Loans

Financial commitments are increased by many loans. When well handled, they demonstrate discipline in repayment, which raises credit scores; when poorly managed, they indicate danger.

#### 16. Debt\_to\_Income\_Ratio

A fundamental indicator of affordability. While high ratios indicate repayment pressure, lower levels indicate income comfortably supports debt commitments.

#### 17. Total\_Debt

A large amount of outstanding debt makes repayment more difficult and raises the financial load. In general, low debt levels are advantageous for evaluating credit risk.

#### 18. EMI

High EMIs in comparison to income limit the flexibility of cash flow. Smaller EMIs carry less risk and are simpler to handle.

#### 19. Loan\_Amount\_Requested

A larger loan request in comparison to income necessitates more robust evidence of repayment capabilities. Requests that are out of scale may raise red flags.

#### 20. Repayment\_Period\_Months

Longer periods raise total interest and long-term exposure to default risk, but they also lower monthly EMIs. Shorter periods have higher monthly pressure but pay off debt more quickly.

#### 21. Credit\_Card\_Utilization\_Rate

Reliance on borrowed cash is indicated by using a large percentage of available credit. A low usage rate (less than 30%) indicates effective credit management.

#### 22. Bank\_Balance

A good bank balance demonstrates liquidity, which instills trust in the ability to repay. Financial strain may be indicated by low balances.

#### 23. Savings\_Balance

Savings serve as a contingency fund. Resilience to changes in income is shown by high savings.

---

24. Credit\_Inquiries\_Last12m

Aggressive borrowing attempts are suggested by several credit checks. This may indicate a risk of overextension or desperation.

25. OnTime\_Payment\_Rate

One of the best indicators of credit is a high rate, which demonstrates constant discipline in fulfilling commitments.

26. Missed\_Payments

Regularly missing payments raises the likelihood of default and negatively impacts credit scores.

27. Mobile\_Data\_Usage\_GB

A borrower's level of digital connectivity may be reflected in their mobile data consumption. High utilization could indicate that e-commerce platforms, banking apps, and online financial services are being used actively. This may be a sign of familiarity with digital payment mechanisms, which are frequently associated with transactions that are traceable and transparent. Low usage could indicate that access to digital banking technologies is restricted.

28. Monthly\_Call\_Minutes

A constant and steady volume of calls may be a subliminal sign of stable living conditions. Individuals with consistent communication styles could also have steady jobs or connections in the community. Such behavioral characteristics serve as a stand-in for dependability and consistency in other domains, such financial habits, in certain risk models.

29. Monthly\_SMS\_Count

Engagement with digital financial services may be indicated by SMS activity, particularly that pertaining to banking alerts and OTPs. A healthy amount of this activity could indicate that finances are being actively monitored. Low or nonexistent SMS activity may indicate a lack of awareness of account activity, which could raise the possibility of fraud or missed payments.

30. No\_of\_Social\_Accounts

Digital literacy and adaptation to online systems, such as digital banking, are sometimes shown by having many social media accounts. With permission, social data may occasionally be utilized to evaluate employment information, location consistency, and stability. The lender may not have access to such other data sources due to low or nonexistent social presence.

31. Avg\_Daily\_Social\_Usage\_Min

While excessive use may be linked to risky lifestyle choices or impulsive purchasing, moderate



---

daily use can suggest healthy social network engagement. Extremely low usage, on the other hand, might indicate less digital engagement, which could limit our understanding of behavioral data. More important than the total minutes is the usage pattern.

32. Facebook\_Friends\_Count

A sizable, steady social network may indicate stability, ties to the community, and potential unofficial financial assistance in times of need. Unexpectedly significant shifts in this number could be a sign of problems in life. These social stability indicators can enhance alternative credit scoring models even if they are not a direct credit factor.

33. Instagram\_Followers\_Count

High follower counts may be a sign of career potential (influencer income streams, for example), which could offer extra, occasionally unusual revenue streams. But it can also mean spending more on lifestyle in order to keep up an online presence. Here, contextual analysis of content kind and engagement is crucial.

34. Monthly\_Ecommerce\_Spend

Spending patterns on the internet can reveal a person's level of income and risk tolerance. Healthy consuming behavior is suggested by consistent but controlled spending, whereas impulsive tendencies may be indicated by irregular or high-risk purchases (such as frequent high-value luxury items). Single-month increases are not as significant as long-term patterns.

35. No\_of\_Transactions\_Per\_Month

An active financial life, consistent income flow, and steady consumption patterns are frequently reflected in a larger transaction count. Predictable cash flow can be indicated by a consistent transaction frequency across several months. Unusual spending patterns or unstable income may be indicated by abrupt decreases or increases.

36. Avg\_Transaction\_Value

Significant transaction frequency combined with large average transaction volumes may indicate active spending, but they can also signal significant income. Frequent but modest transactions frequently indicate that daily costs are being well controlled. When determining affordability, the transaction value to income ratio is crucial.

37. Subscription\_Service\_Count

Monthly commitments are reflected in recurring subscriptions, including gym memberships or streaming services. An excessive number of subscriptions might lower discretionary income,

---

particularly if income is tight. In general, a moderate number of subscriptions in relation to income presents little danger of repayment.

38. Installment\_Purchase\_Count

Using credit to pay for lifestyle needs may be indicated by multiple installment purchases. Excessive counts might put a burden on monthly budgets, but affordable installments can spread expenditures. An effective indicator of overall loan repayment capacity is the tracking of payback trends for these purchases.

39. Avg\_UPI\_Transactions

Strong digital payment uptake and open spending habits are indicated by high UPI activity. Additionally, it may show regular but controllable transactions, indicating sound financial management. Extremely low UPI usage may indicate a greater reliance on cash by the borrower, which would limit data for risk models.

40. UPI\_Count

The amount of transaction activity can be inferred from the monthly number of UPI transactions. Regular financial engagement is demonstrated by a consistent, reasonable count. While extremely low counts could suggest cash reliance or financial inactivity, extremely high levels might imply extensive spending behaviors.

41. Avg\_Monthly\_UPI\_Transaction\_Value

High income levels may be associated with higher average transaction values, particularly when combined with regular UPI use. Large, erratic payments, however, can indicate erratic spending. Low numbers may indicate steady consumption patterns and frequently represent daily payments.

42. Utility\_Bill\_Payment\_Timeliness

One of the best behavioral indicators of creditworthiness is timely bill payment. On-time payments demonstrate discipline and dependability in fulfilling commitments. Similar loan payback delays may be predicted by regular utility bill payment delays.

43. SIM\_Swap\_History

Since scammers occasionally switch SIMs in order to intercept OTPs, frequent SIM switches may be a warning sign for possible fraud. Better trustworthiness and reduced identity risk are indicated by a steady SIM history. This could be used by lenders to identify potential high-risk borrowers.

44. No\_of\_Devices\_Linked

A tech-savvy borrower with active digital involvement may have multiple devices connected to

---

their bank accounts. Although generally beneficial, if the account holder is negligent with device management, having too many devices could cause security issues. Consistency and regulated access are suggested by stable device linkage patterns.

#### 45. Credit\_Score

A person's creditworthiness is indicated by their credit score, which in various scoring methods ranges from 300 to 850. Payment history, credit use, credit history duration, credit account kinds, and recent queries are all taken into consideration while calculating it. Lenders view a higher score as a sign of less risk, which opens the door to better lending conditions and cheaper interest rates. Higher default risk is indicated by a low score, which frequently results in more stringent borrowing requirements..

#### 46. Probability\_Not\_Default

This indicator shows how likely it is that a borrower will fulfill all of their repayment commitments without going into default, shown as a percentage or probability value. It frequently comes from statistical or machine learning models that evaluate a variety of behavioral and financial aspects. Very low default risk is indicated by a value near 1 (or 100%), whereas significant risk is suggested by a value around 0. Lenders adjust approval decisions and interest rates based on this probability.

---

# Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is the act of looking at and visualizing data in order to learn about its structure, patterns, and important relationships prior to model building. It entails summarizing statistical quantities and producing graphics to identify trends, anomalies, and correlations. EDA would ensure that subsequent analysis is based on a solid understanding of the data and serves to inform feature selection, preprocessing operations, and model selection.

We have performed the following steps on our dataset in order to know more about the dataset.

## 5.1 Descriptive Statistics

	Age	Gender	Marital_Status	Education_Level	Employment_Status	Occupation	Annual_Income	Monthly_Income	Household_Size
count	200000.000000	200000.000000	200000.000000	200000.000000	200000.000000	200000.000000	200000.000000	200000.000000	200000.000000
mean	40.271065	0.370430	0.443795	0.449345	1.497220	2.497020	56605.864320	4717.155377	2.172715
std	11.331950	0.482921	0.830984	0.816246	1.119084	1.708346	29595.232728	2466.269452	0.907661
min	21.000000	0.000000	0.000000	0.000000	0.000000	0.000000	10000.000000	833.330000	1.000000
25%	32.000000	0.000000	0.000000	0.000000	0.000000	1.000000	35728.505000	2977.375000	2.000000
50%	40.000000	0.000000	0.000000	0.000000	1.000000	2.000000	50071.555000	4172.630000	2.000000
75%	48.000000	1.000000	1.000000	0.000000	2.000000	4.000000	70139.135000	5844.930000	3.000000
max	69.000000	1.000000	3.000000	2.000000	3.000000	5.000000	200000.000000	16666.670000	5.000000

Figure 5.1

Dependents	...	Installment_Purchase_Count	Avg_UPI_Transactions	UPI_Count	Avg_Monthly_UPI_Transaction_Value	Utility_Bill_Payment_Timeliness
200000.000000	...	200000.000000	200000.000000	200000.000000	200000.000000	200000.000000
0.621865	...	4.507500	25.019775	5.004900	580.927882	0.727081
1.033162	...	2.872554	9.926631	2.230417	402.750706	0.174450
0.000000	...	0.000000	0.000000	0.000000	34.260000	0.060000
0.000000	...	2.000000	18.000000	3.000000	313.250000	0.610000
0.000000	...	5.000000	25.000000	5.000000	477.490000	0.760000
1.000000	...	7.000000	32.000000	6.000000	726.005000	0.870000
3.000000	...	9.000000	66.000000	17.000000	6526.840000	1.000000

Figure 5.2

---

---

SIM_Swap_History	No_of_Devices_Linked	Geo_Location_Variance_Score	Credit_Score	Probability_Not_Default
200000.000000	200000.000000	200000.000000	200000.000000	200000.000000
0.998135	2.497570	0.501155	388.912625	0.055379
0.817389	1.117019	0.223523	58.009181	0.048116
0.000000	1.000000	0.000000	300.000000	0.011000
0.000000	1.000000	0.330000	343.000000	0.021000
1.000000	2.000000	0.500000	389.000000	0.041000
2.000000	3.000000	0.680000	431.000000	0.073000
2.000000	4.000000	1.000000	619.000000	0.571000

Figure 5.3

## 5.2 Missing Value Analysis

We have checked that whether any value is missing or not in our dataset.

```
1 data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200000 entries, 0 to 199999
Data columns (total 46 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Age                                   200000 non-null float64
1   Gender                               200000 non-null int64
2   Marital_Status                       200000 non-null int64
3   Education_Level                      200000 non-null int64
4   Employment_Status                   200000 non-null int64
5   Occupation                           200000 non-null int64
6   Annual_Income                        200000 non-null float64
7   Monthly_Income                       200000 non-null float64
8   Household_Size                       200000 non-null int64
9   Dependents                           200000 non-null int64
10  Home_Ownership                       200000 non-null int64
11  Residence_Type                       200000 non-null int64
12  Credit_History_Length                 200000 non-null int64
13  Number_of_Credit_Cards                200000 non-null int64
14  Number_of_Loans                       200000 non-null int64
15  Debt_to_Income_Ratio                  200000 non-null float64
16  Total_Debt                           200000 non-null float64
17  EMI                                   200000 non-null float64
18  Loan_Amount_Requested                 200000 non-null float64
19  Repayment_Period_Months               200000 non-null int64
20  Credit_Card_Utilization_Rate           200000 non-null float64
21  Bank_Balance                          200000 non-null float64
```

Figure 5.4

```
20  Credit_Card_Utilization_Rate           200000 non-null float64
21  Bank_Balance                          200000 non-null float64
22  Savings_Balance                       200000 non-null float64
23  Credit_Inquiries_Last12m              200000 non-null int64
24  OnTime_Payment_Rate                   200000 non-null float64
25  Missed_Payments                       200000 non-null int64
26  Mobile_Data_Usage_GB                  200000 non-null float64
27  Monthly_Call_Minutes                   200000 non-null float64
28  Monthly_SMS_Count                     200000 non-null int64
29  No_of_Social_Accounts                  200000 non-null int64
30  Avg_Daily_Social_Usage_Min             200000 non-null float64
31  Facebook_Friends_Count                 200000 non-null int64
32  Instagram_Followers_Count              200000 non-null int64
33  Monthly_Ecommerce_Spend                200000 non-null float64
34  No_of_Transactions_Per_Month           200000 non-null int64
35  Avg_Transaction_Value                  200000 non-null float64
36  Subscription_Service_Count              200000 non-null int64
37  Installment_Purchase_Count              200000 non-null int64
38  Avg_UPI_Transactions                   200000 non-null float64
39  UPI_Count                             200000 non-null int64
40  Avg_Monthly_UPI_Transaction_Value       200000 non-null float64
41  Utility_Bill_Payment_Timeliness         200000 non-null float64
42  SIM_Swap_History                       200000 non-null int64
43  No_of_Devices_Linked                   200000 non-null int64
44  Credit_Score                           200000 non-null float64
45  Probability_Not_Default                 200000 non-null float64
dtypes: float64(21), int64(25)
memory usage: 70.2 MB
```

Figure 5.5

## 5.3 Correlation Analysis

We have also checked that that all the features are linearly independent or not by checking correlation between them. So for this we plotted the corr matrix using heat-map .

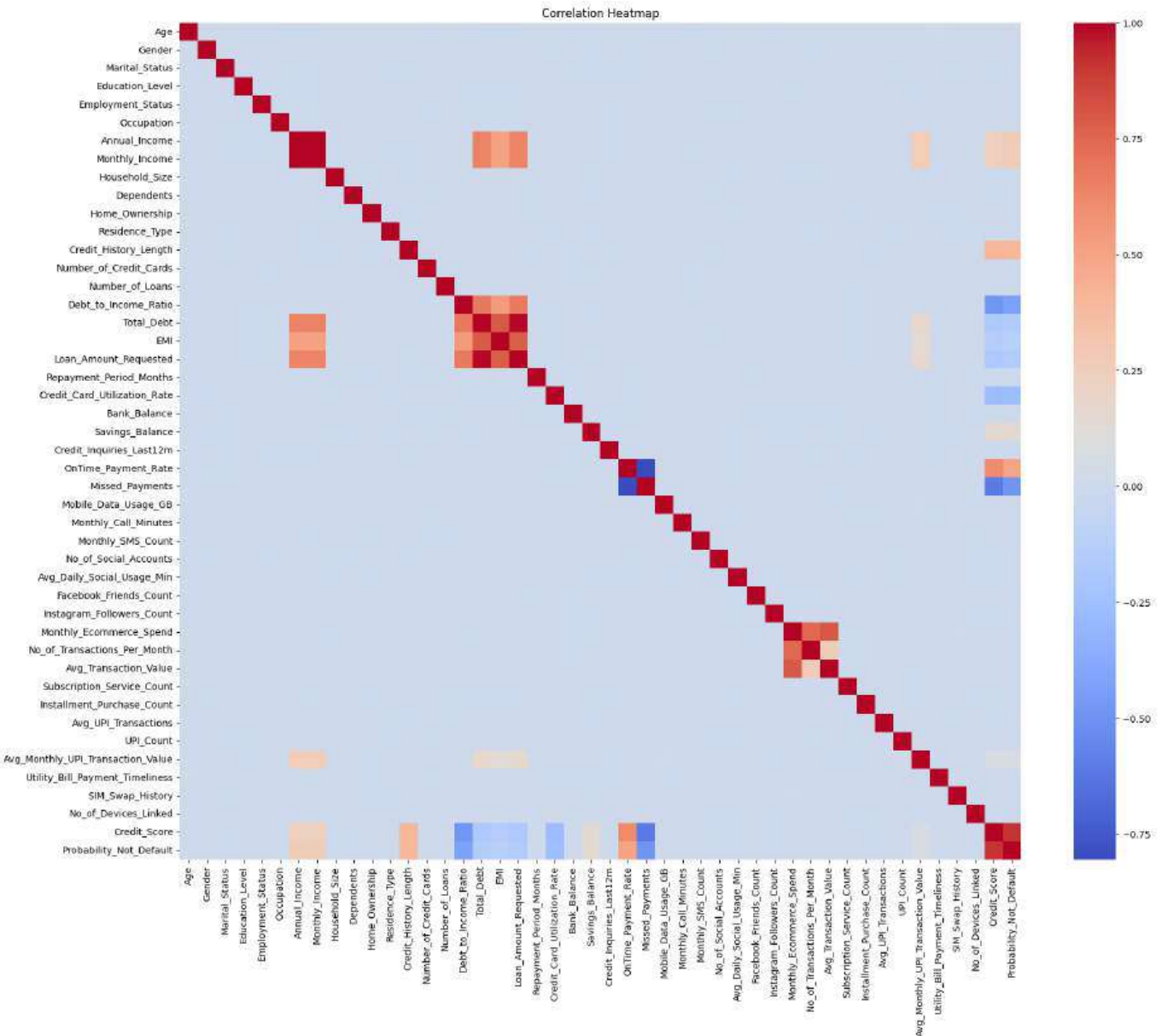


Figure 5.6

---

## 5.4 Insights from EDA

After doing Exploratory Data Analysis on our data we got the better understanding of our data, what does it contains. On doing EDA we found that data was not uniformly scaled. For eg. Age is in order of 21 to 69 where as OnTime\_Payment\_Rate is in order of 0.01 to 1 so we have to scale all the data normally so that weights will be in same order and model can be train properly. Another thing which we found that some of the columns are linearly correlated with each other so we have to drop or create the appropriate features from them and try to reduce the correlations between the features. So that part will be covered in feature engineering part.













---

# Feature Engineering

## 6.1 Encoding Categorical Variables

We have many categorical variables in our dataset therefore in order to use it for regression we have encoded each categorical value as follow using the previous work done for alternative data. Now the reference for each encoding is shown below:

Feature Name	Description	Encoding Method Used	Source(s)
Gender	Borrower's gender	Integer mapping	Kumar & Babu (2025) 
Marital_Status	Marital status category	Integer mapping / One-Hot	Kumar & Babu (2025) 
Education_Level	Highest educational qualification	Ordinal mapping	Kumar & Babu (2025)  ; Alamsyah et al. (2025) 
Employment_Status	Employment type	One-Hot	Kumar & Babu (2025) 
Occupation	Job category / industry type	Target encoding / One-Hot	Kumar & Babu (2025)  ; Alamsyah et al. (2025) 
Home_Ownership	Home ownership status	One-Hot	Kumar & Babu (2025) 
Residence_Type	Type of residence	One-Hot	Kumar & Babu (2025) 
SIM_Swap_History	Number of SIM swaps (fraud risk indicator)	Integer mapping	Niu et al. (2019) 




Figure 6.1

## 6.2 Scaling and Normalization

As we have already seen in EDA section that the features are on different scale so we have used normalization for each feature. For this we have used z score normalization technique

The Z score normalization can be done as follow:

$$z = \frac{x - \mu}{\sigma}$$

Figure 6.2

---

where,

$Z$  = new normalized value

$X$  = original sample

$\mu$  = mean of the population

$\sigma$  = standard deviation of the population

By this we were able to scale over data in same range so that values of weight will small and model training will be more accurate.

## 6.3 Dimensionality Reduction

During EDA on over dataset we show that the some of the columns are linearly related to the other column so we dropped some features as they are having same meaning or was indicating to same thing . We have kept the threshold of 0.7 for correlation to see which columns are correlating with each other.

So that correlating features are as follow;

1. Monthly\_Income and Annual\_Income: 1.0000
2. Loan\_Amount\_Requested and Total\_Debt: 0.9888
3. Missed\_Payments and OnTime\_Payment\_Rate: -0.8052
4. Avg\_Transaction\_Value and Monthly\_Ecommerce\_Spend: 0.7984
5. EMI and Total\_Debt: 0.7927
6. Loan\_Amount\_Requested and EMI: 0.7838
7. No\_of\_Transactions\_Per\_Month and Monthly\_Ecommerce\_Spend: 0.7408

So now we removed Monthly\_Income , No\_of\_Transactions\_Per\_Month , Avg\_Transaction\_Value and Missed\_Payments as they are having the same meaning.

---

# Model Development

Now After doing EDA and feature engineering on the dataset we have split the dataset in ratio 4:1 . So we kept 80 % data for training and 20 % data for the testing purpose. We have used different different model for training purpose

## 7.1 Problem Formulation – Multi-Output Regression for PD & Credit Score

The goal of this project is to design and implement a predictive modeling framework capable of simultaneously estimating two continuous credit risk–related targets:

1. Credit Score: A numerical indicator of a borrower’s creditworthiness, traditionally ranging between 300 and 850 in standard scales.
2. Probability of Not Being a Defaulter (PD\_Not\_Default): A probability value between 0 and 1, representing the model’s confidence that a customer will not default within the defined risk horizon.

Rather than building two independent models, this task is formulated as a multi-output regression problem. In this setup, a single predictive model learns a shared representation from the same set of features and outputs both target variables in parallel. This approach has the following advantages:

- ◆ Shared Learning Across Targets – Many of the underlying predictors (e.g., annual income, on-time payment ratio, existing debt, ecommerce spending behavior) influence both credit score and default probability. Multi-output regression allows the model to capture and exploit these correlations.
- ◆ Consistent Predictions – Using one model for both outputs ensures that the predicted credit score and the probability of non-default remain aligned logically (e.g., a high credit score is unlikely to correspond to a low probability of repayment).
- ◆ Efficiency – Fewer models to train, maintain, and deploy, reducing computational and operational costs.

---

## Mathematical Formulation

Let:

$X \in \mathbb{R}^{n \times p}$  be the matrix of  $n$  samples and  $p$  features (in our case, 45 engineered and encoded predictors).

$Y \in \mathbb{R}^{n \times 2}$  be the target matrix, where:

$y_{:,1}$  = Credit Score (continuous)

$y_{:,2}$  = Probability of Not Default (continuous in  $[0,1]$ )

The objective is to learn a function:

$$f : \mathbb{R}^p \rightarrow \mathbb{R}^2$$

Figure 7.1

that minimizes a joint loss over both targets, typically the Mean Squared Error (MSE):

$$\mathcal{L}(f) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^2 (y_{ij} - \hat{y}_{ij})^2$$

Figure 7.2

## Machine Learning and Deep Learning Approaches

To solve this, a diverse set of supervised regression algorithms were applied, ranging from classical ML models to advanced neural networks:

### Classical ML

- i. Linear Regression (baseline)
- ii. Random Forest Regressor
- iii. XGBoost Regressor
- iv. LightGBM Regressor

### Deep Learning

- 
- I. Multi-Layer Perceptron (MLP) for tabular data
  - II. 1D Convolutional Neural Network (CNN) for sequential feature patterns
  - III. Long Short-Term Memory (LSTM) networks for temporal/sequence modeling

For neural networks, the architecture is defined with a shared hidden representation and a final dense output layer of size 2 (one for each target). The loss function is the combined MSE for both outputs.

### **Why Multi-Output Regression Fits This Problem??**

The strong correlation between credit score and probability of non-default makes a joint learning approach more data-efficient and coherent than building isolated models. In practice, this improves generalization performance and ensures that business decisions derived from both predictions are mutually consistent.

## **7.2 Baseline Model Selection (Linear Regression)**

As a first step, a baseline model was established to serve as a performance benchmark for all subsequent experiments. For this, a multi-output linear regression model was chosen due to its simplicity, interpretability, and minimal computational cost.

### **Rationale for Choice:**

- ◆ Interpretability – The model coefficients directly indicate the direction and magnitude of each feature's influence on the outputs (Credit Score and Probability of Not Default)
- ◆ Speed – Training is instantaneous even on large datasets.
- ◆ Baseline Benchmark – Provides a lower bound for expected performance; any advanced model must outperform this in both RMSE and  $R^2$ .

### **Model Characteristics:**

- ◆ Fitted independently for each target variable but within a multi-output regression wrapper.
- ◆ Assumes linear relationships between predictors and targets.
- ◆ No regularization applied initially, to preserve interpretability.

---

### **Performance Summary:**

While the linear regression model performed adequately in capturing broad trends, it failed to account for the non-linear feature-target relationships present in the data. This was evident in lower  $R^2$  scores and relatively high RMSE values, particularly for the Probability of Not Default, which exhibited more complex patterns than linear assumptions could capture.

### **Random Forest Regressor**

The Random Forest Regressor (RFR) was selected as the first non-linear benchmark model due to its robustness, interpretability, and strong performance in tabular datasets.

#### **Key Properties:**

1. Ensemble Method: Aggregates predictions from multiple decision trees, reducing variance compared to a single tree.
2. Non-Parametric: Captures complex, non-linear relationships without assuming functional form.
3. Feature Importance: Provides insights into which variables have the greatest influence on predictions.

#### **Model Configuration:**

- ❖ Implemented via `sklearn.ensemble.RandomForestRegressor` in multi-output mode.
- ❖ Number of trees (`n_estimators`) tuned in the range of 100–500.
- ❖ Maximum tree depth (`max_depth`) optimized to balance overfitting and generalization.
- ❖ Minimum samples per leaf (`min_samples_leaf`) adjusted to improve stability.
- ❖ Random seed fixed for reproducibility.

#### **Advantages for This Task:**

- ❖ Handles mixed feature types (numeric and encoded categorical) naturally.
- ❖ Resistant to outliers due to median-based tree splits.
- ❖ Well-suited to high-dimensional data (45 features) without requiring extensive scaling.

---

### **Performance Summary:**

The Random Forest Regressor substantially outperformed the baseline linear regression in both RMSE and  $R^2$  for Credit Score and Probability of Not Default. It captured interaction effects between behavioral, financial, and demographic variables that the baseline missed. Additionally, the feature importance analysis revealed that variables such as Annual Income, On-Time Payment Ratio, and Existing Debt had the strongest influence across both outputs.

### **Gradient Boosting Models (XGBoost & LightGBM)**

Gradient Boosting Machines (GBMs) are an advanced ensemble learning technique that sequentially builds decision trees, with each new tree correcting the residual errors of the ensemble's previous predictions. This approach often delivers superior accuracy over bagging methods (e.g., Random Forest) due to its iterative error minimization process.

In this project, two state-of-the-art GBM implementations were evaluated:

1. XGBoost Regressor: Highly optimized gradient boosting with regularization and parallel processing.
2. LightGBM Regressor: Gradient boosting optimized for speed and memory efficiency, particularly effective on large-scale and high-dimensional datasets.

### **Why GBMs for Credit Risk Prediction**

GBMs excel at:

- ❖ Capturing Non-Linear Relationships: Particularly important given the complex interplay between behavioral, transactional, and demographic predictors in credit risk modeling.
- ❖ Handling Mixed Feature Types: Works seamlessly with numeric and encoded categorical features.
- ❖ Feature Interaction Learning: Automatically discovers and models high-order feature interactions without manual engineering.
- ❖ Imbalanced Sensitivity: Handles skewed target distributions better than many classical algorithms, useful when default-related behaviors are rare.

---

## Model Configurations

### ❖ XGBoost Regressor (xgboost.XGBRegressor)

n\_estimators: 200–1000

learning\_rate: 0.01–0.3

max\_depth: 3–10

subsample & colsample\_bytree: 0.6–1.0

reg\_alpha & reg\_lambda: Regularization terms tuned to reduce overfitting.

### ❖ LightGBM Regressor (lightgbm.LGBMRegressor)

n\_estimators: 200–1000

learning\_rate: 0.01–0.3

max\_depth: -1 (unlimited) or restricted to 5–15 for regularization

num\_leaves: 31–255

feature\_fraction & bagging\_fraction: 0.6–1.0

lambda\_l1 & lambda\_l2: Regularization parameters.

Both models were run in multi-output mode by training separate but identically tuned regressors for each target variable (Credit Score and Probability of Not Default). This was necessary because native multi-output regression is not directly supported in XGBoost/LightGBM; instead, the problem was handled via a wrapper approach in our training pipeline.

## Performance Summary compared to Random Forest:

❖ **Higher Predictive Accuracy:** Both XGBoost and LightGBM achieved lower RMSE and higher  $R^2$  for both targets, with LightGBM slightly outperforming XGBoost in terms of computation time. Better Generalization – GBMs maintained strong performance on the test set, indicating effective overfitting control.

❖ **Feature Importance Consistency:** Both models confirmed the top predictors identified by Random Forest (e.g., Annual Income, On-Time Payment Ratio, Existing Debt) while also surfacing additional nuanced variables such as Monthly eCommerce Spend and Credit Card Usage Ratio.



---

## Deep Learning Models (MLP, CNN, LSTM)

While tree-based ensemble methods such as Random Forest, XGBoost, and LightGBM offer exceptional performance for tabular data, deep learning models can sometimes surpass them when provided with large datasets, appropriate feature scaling, and well-tuned architectures. For this project, three deep learning architectures were evaluated for multi-output regression:

1. Multi-Layer Perceptron (MLP): Fully connected feed-forward network for general-purpose tabular learning.
2. Convolutional Neural Network (1D-CNN): Convolutional layers adapted for sequential or locally correlated feature patterns.
3. Long Short-Term Memory (LSTM): Recurrent neural network architecture designed to model temporal or ordered feature dependencies.

### 1. Multi-Layer Perceptron (MLP)

Architecture:

Input layer matching the 45 encoded features.

2–4 hidden layers with 32–128 neurons each.

Dropout layers (0.1–0.5) to mitigate overfitting.

ReLU activation for hidden layers, linear activation for the 2-output layer.

Advantages:

Learns complex, non-linear feature-target relationships.

Straightforward to implement and tune with scikeras and RandomizedSearchCV.

Outputs both Credit Score and Probability of Not Default simultaneously.

### 2. 1D Convolutional Neural Network (CNN)

Architecture:

Input reshaped to a 1D sequence of features.

One or more convolutional layers (kernel sizes 2–4) with 32–64 filters.

Max-pooling layers to downsample and capture local feature patterns.

---

Dense layers for final regression outputs.

**Advantages:**

Effective at extracting local feature dependencies.

Good at capturing small-scale interactions without explicit feature engineering.

Faster training compared to LSTMs while maintaining competitive performance.

**3. Long Short-Term Memory (LSTM)**

**Architecture:**

- Input treated as a sequence (time-steps = features).
- One or two LSTM layers (32–64 units) to capture sequential patterns in feature interactions.
- Dropout and recurrent dropout applied for regularization
- Dense layers to produce the two regression outputs.

**Advantages:**

- Retains long-range dependencies between features.
- Useful when features have a logical temporal or ordered structure (e.g., historical transaction series, time-based spending behaviors).

**Performance Summary**

❖ MLP consistently outperformed CNN and LSTM on this purely tabular dataset, achieving lower RMSE and higher  $R^2$  on both targets.

❖ CNN was competitive but slightly less accurate, indicating that local convolutional filters did not yield significant benefits for non-sequential credit risk features.

❖ LSTM underperformed compared to MLP and CNN, likely because the dataset lacked true temporal sequences that would leverage its memory capabilities.

---

# Model Evaluation

## 8.1 Evaluation Metrics (MAE, RMSE, R<sup>2</sup>)

To assess model performance, three complementary metrics were employed for each target variable (Credit Score and Probability of Not Default):

### 1. Mean Absolute Error (MAE):

Mean Absolute error (MAE) is the average absolute difference between the predicted values ( $\hat{y}$ ) and the actual values ( $y$ ). It measures the average magnitude of prediction errors, without considering their direction. It is less sensitive to large outliers and it is also interpretable in the same units as the target variable.

MAE can be calculated using following formula.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Figure 8.1

### 2. Root Mean Square error (RMSE):

Root Mean Square Error (RMSE) is the square root of the average of the squared differences between predicted and actual values. It represents the standard deviation of the prediction errors. In this larger errors are penalized more heavily due to the squaring term. Lower values of RMSE indicate better predictive accuracy. RMSE can be calculated as ,

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Figure 8.2

---

### 3. Coefficient of Determination ( $R^2$ Score):

$R^2$  score measures the proportion of variance in the target variable explained by the model. High  $R^2$  score values indicate that the model explains most of the variability in the target variable. It measures how well the model explains variability and gives us a sense of explanatory power.  $R^2$  score can be calculated as ,

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Figure 8.3

Where:

$\bar{y}$  = mean of actual values

$R^2 = 1$  means perfect prediction

$R^2 = 0$  means the model performs no better than predicting the mean value for all observations

---

## 8.2 Residual Analysis

We have done residual analysis to ensure that the models met the assumptions of stability and unbiased. For this we examine the scatter plots of residuals vs. predicted values for each model and for each target variable separately. Ideally, residuals should be randomly dispersed around zero, indicating no systematic error. Here are the model wise plots for each model for both the target variable.

### ❖ Linear Regression

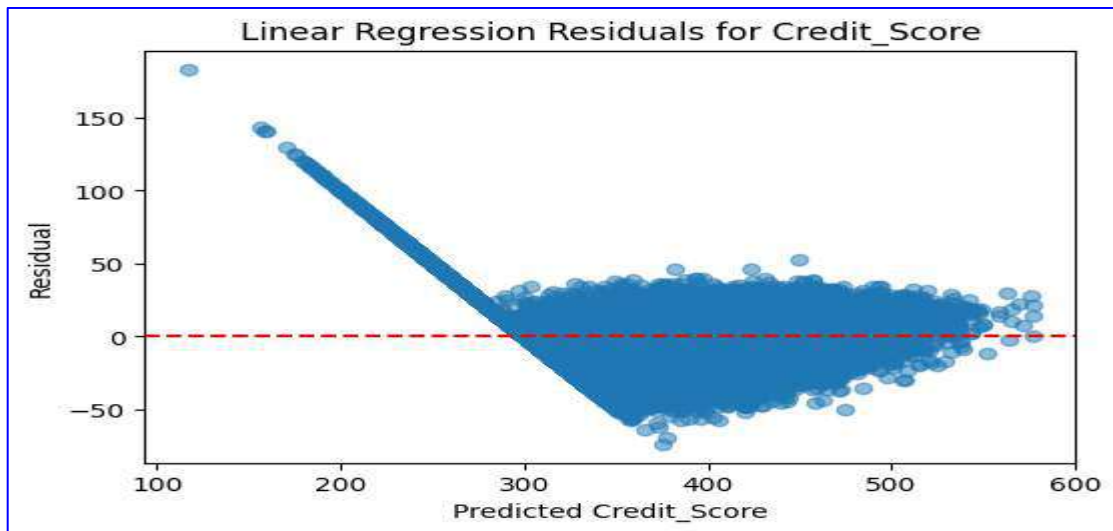


Figure 8.4

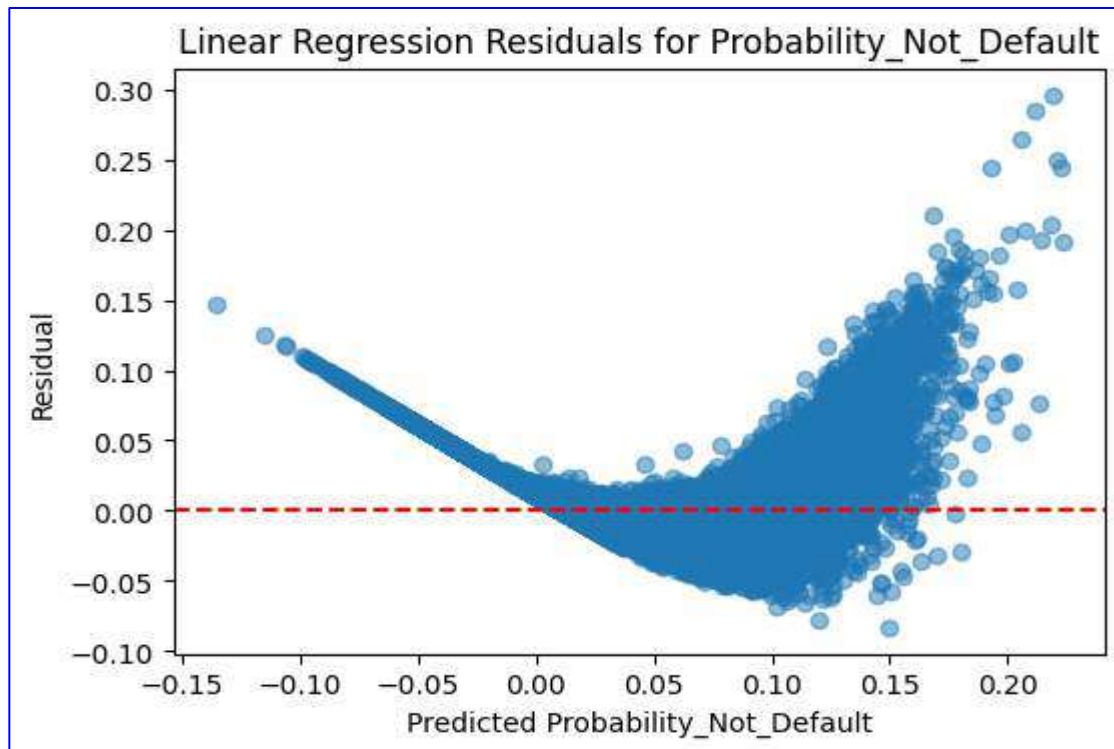


Figure 8.5

As we can see that the residual plot is not symmetric around the  $Y=0$  line we can say that this model is failed to understand the pattern in the data is underfitted .

#### ❖ Random Forest Regression

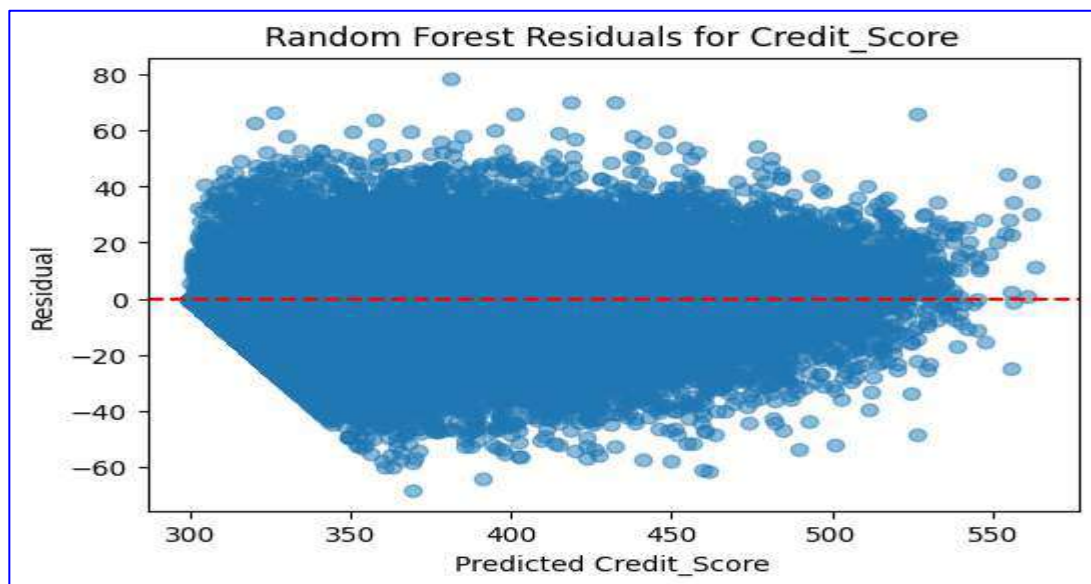


Figure 8.6

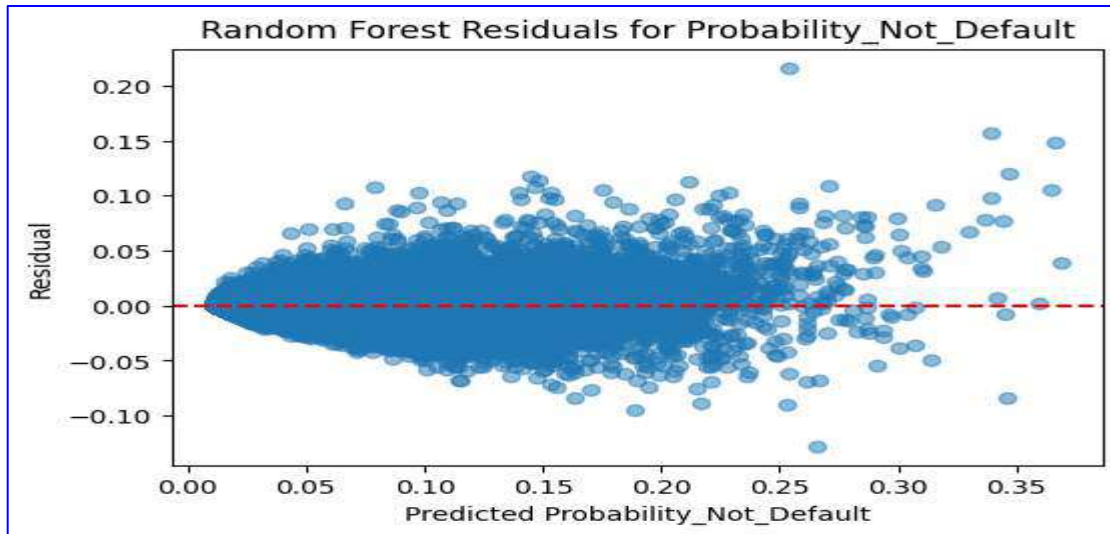


Figure 8.7

This plot for the Random Forest model shows residuals scattered randomly around zero without any clear pattern. It indicates that the model is able to capture the underlying relationship between predictors and target variables. The spread of residuals appears relatively consistent across the range of predicted values. This suggests no major signs of heteroscedasticity. So it indicates that the Random Forest model provides a better fit compared to the linear regression model.

#### ❖ XGBoost Regression

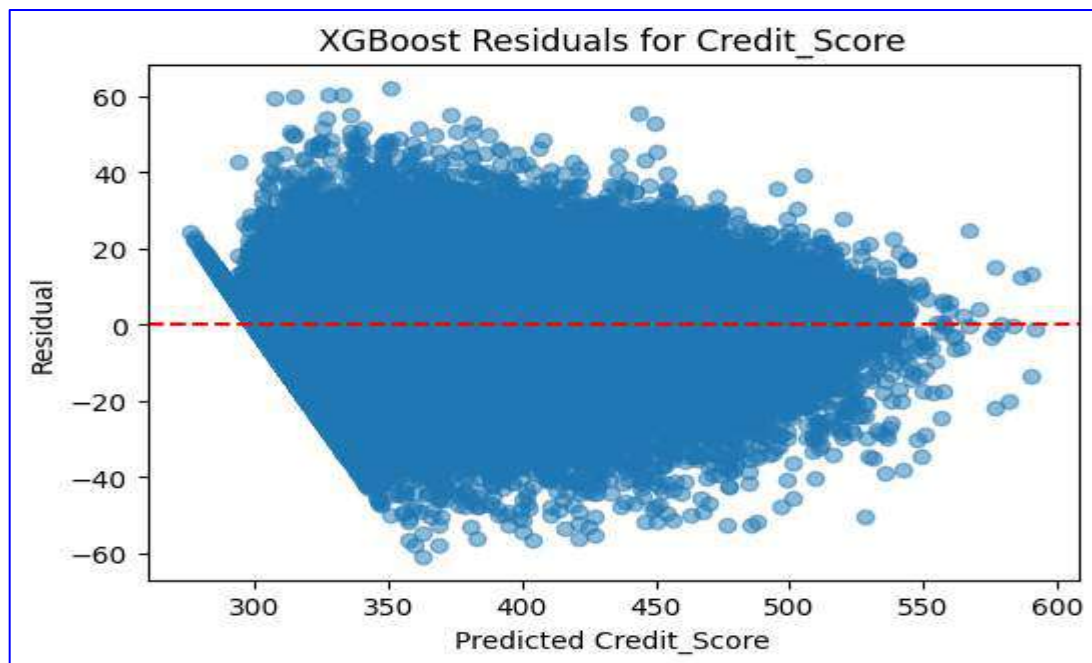


Figure 8.8

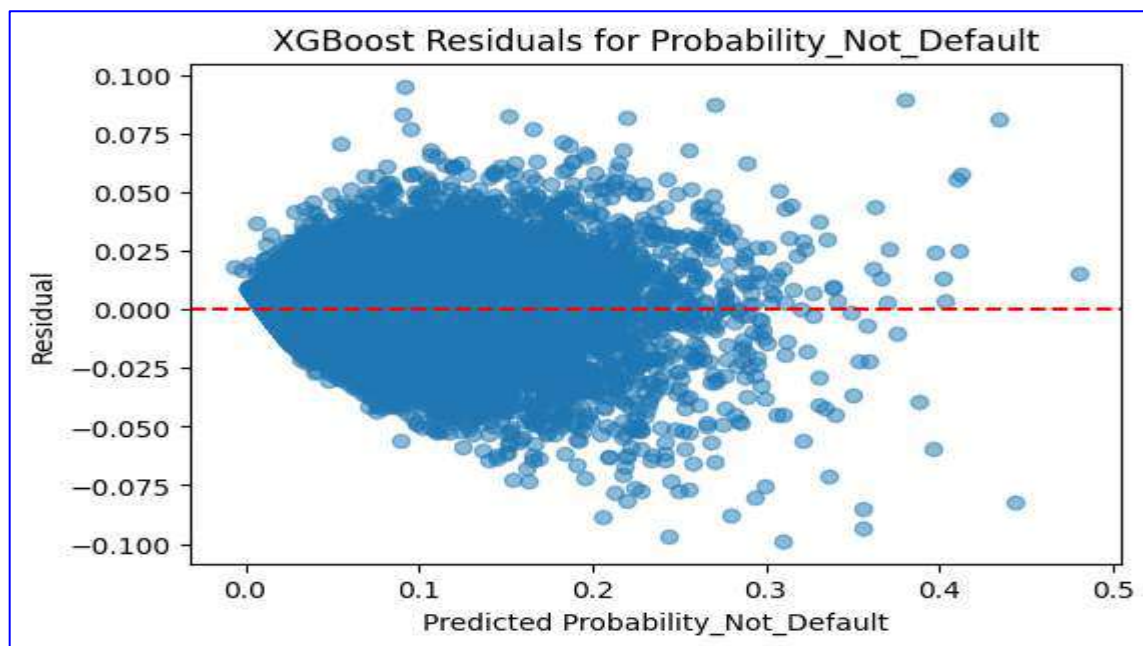


Figure 8.9

❖ LightGBM Regression

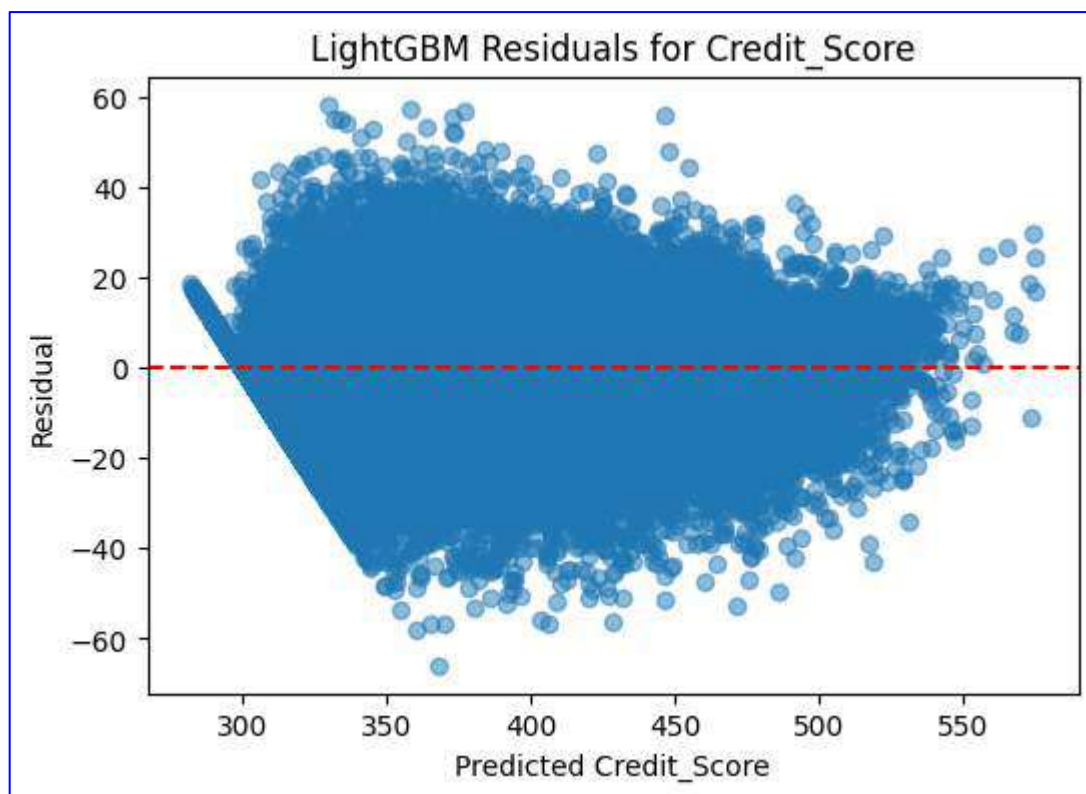


Figure 8.10



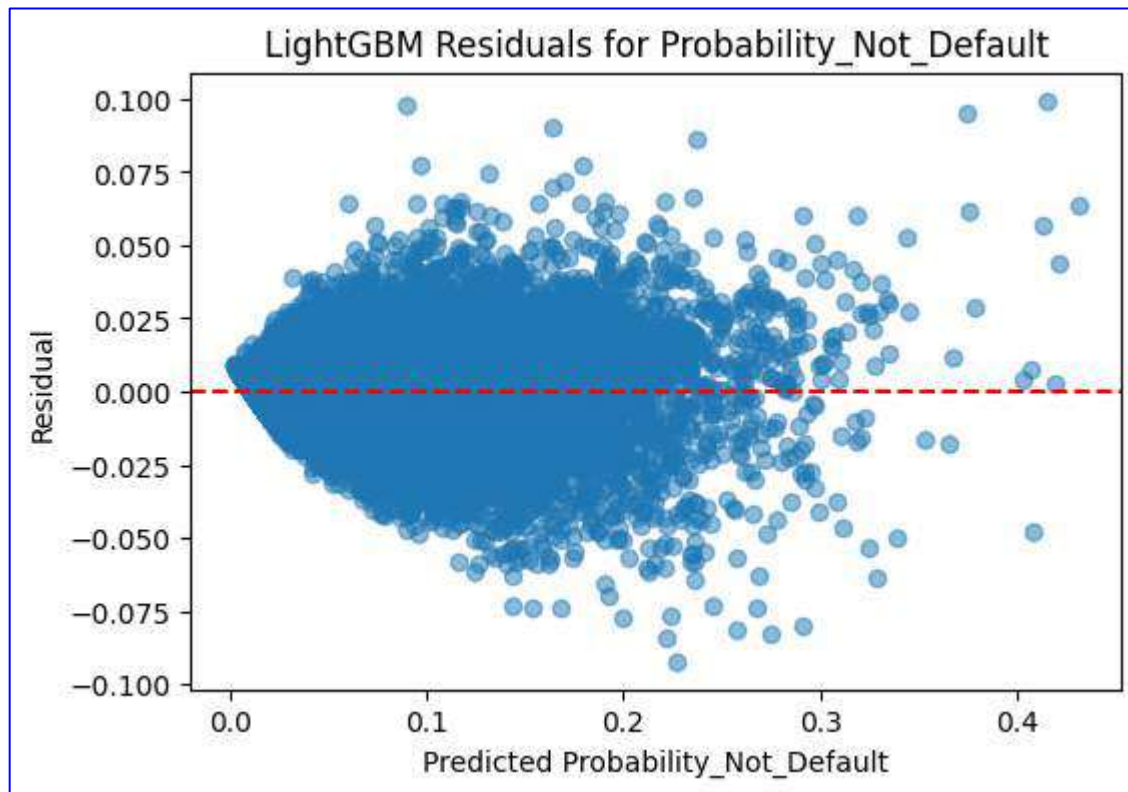


Figure 8.11

These plots for models show residuals scattered randomly around zero without any clear pattern. It indicates that each model is able to capture the underlying relationship between predictors and target variables. The spread of residuals appears relatively consistent across the range of predicted values. This suggests no major signs of heteroscedasticity. So it indicates that each model provides a better fit compared to the linear regression model.

On the basis of this plot analysis we can say that the plot for Random forest, XGBoost and LightGBM are going towards the line  $Y=0$  over time and there is no particular pattern visible for the plot and data is scattered randomly so we can say that the model is learning the relationship between features and target variable. The difference between actual value and predicted output is decreasing gradually, which indicates that the model is learning the pattern between the data.

---

## 8.3 Predicted vs Actual Plots

Here are the model wise predicted vs actual plot for both the target variables.

### ❖ Linear Regression

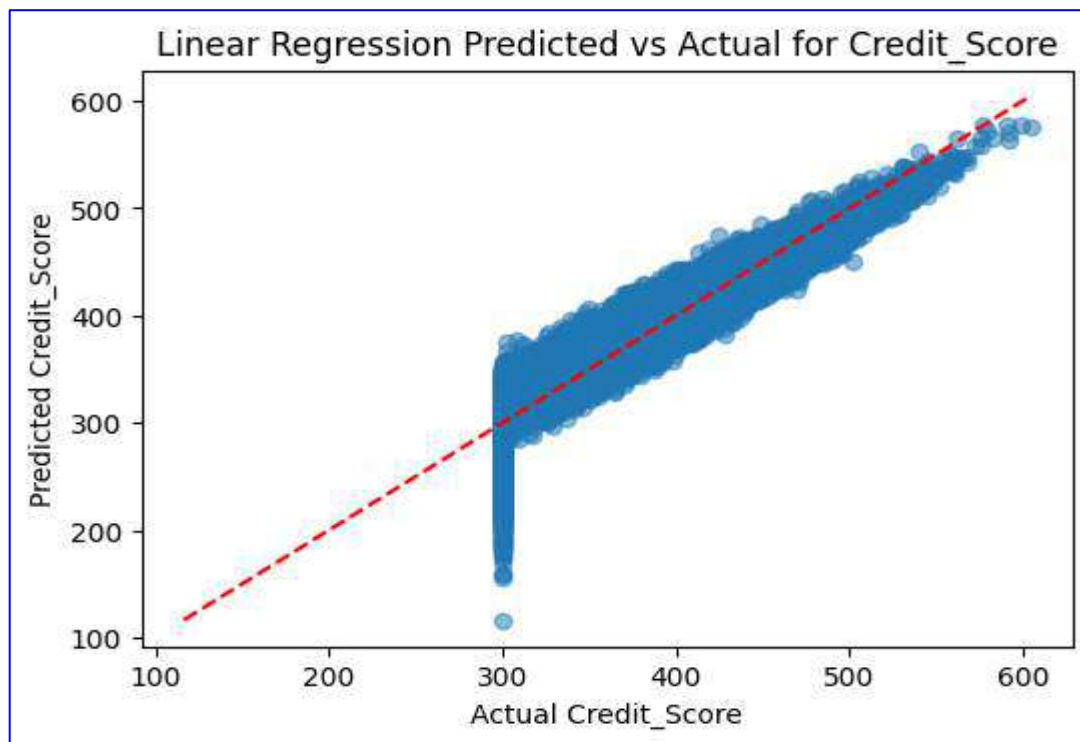


Figure 8.12

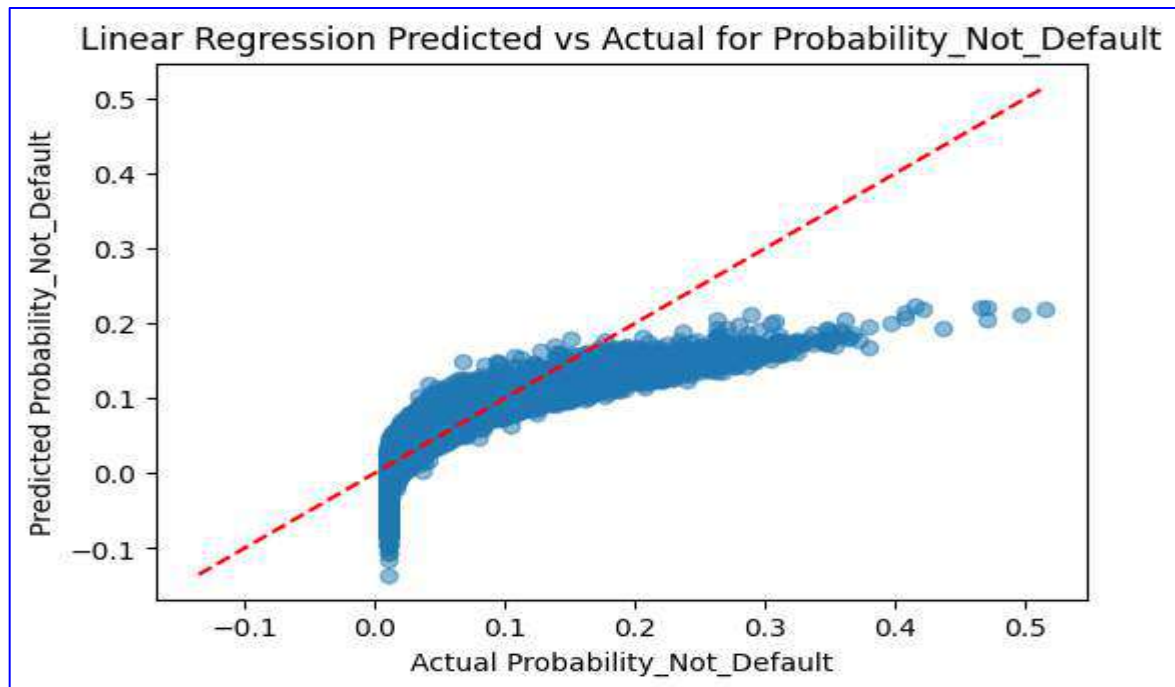


Figure 8.13

❖ **Random Forest**

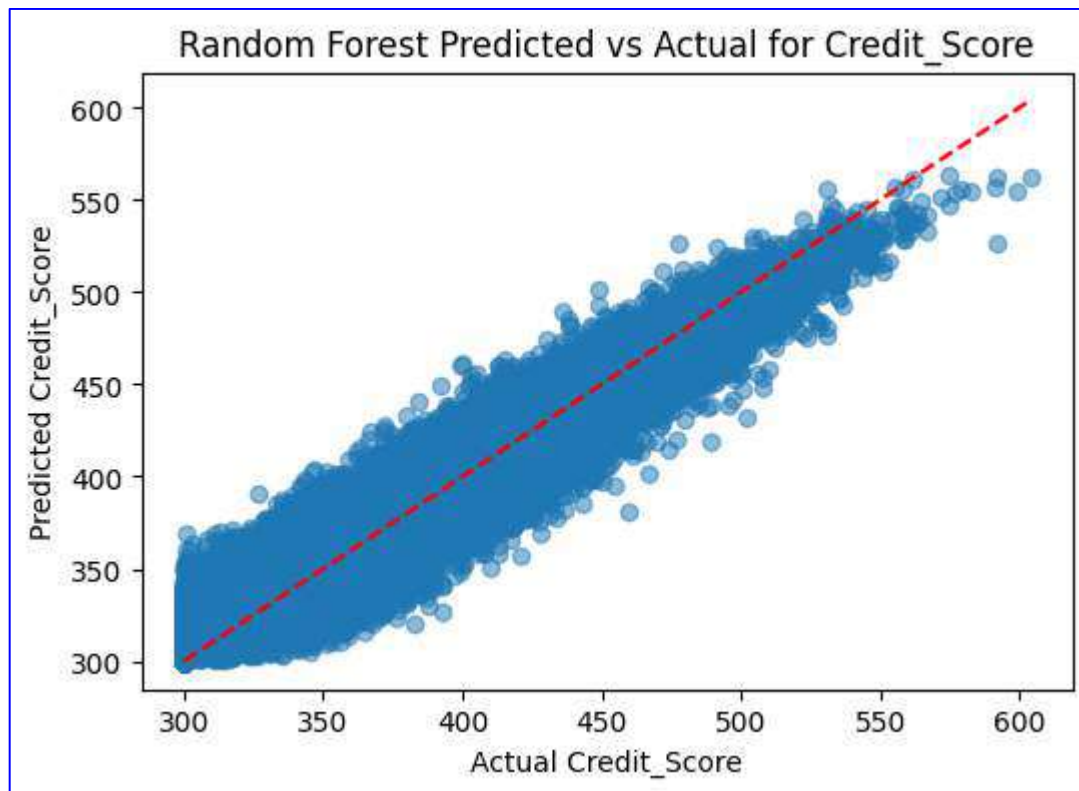


Figure 8.14

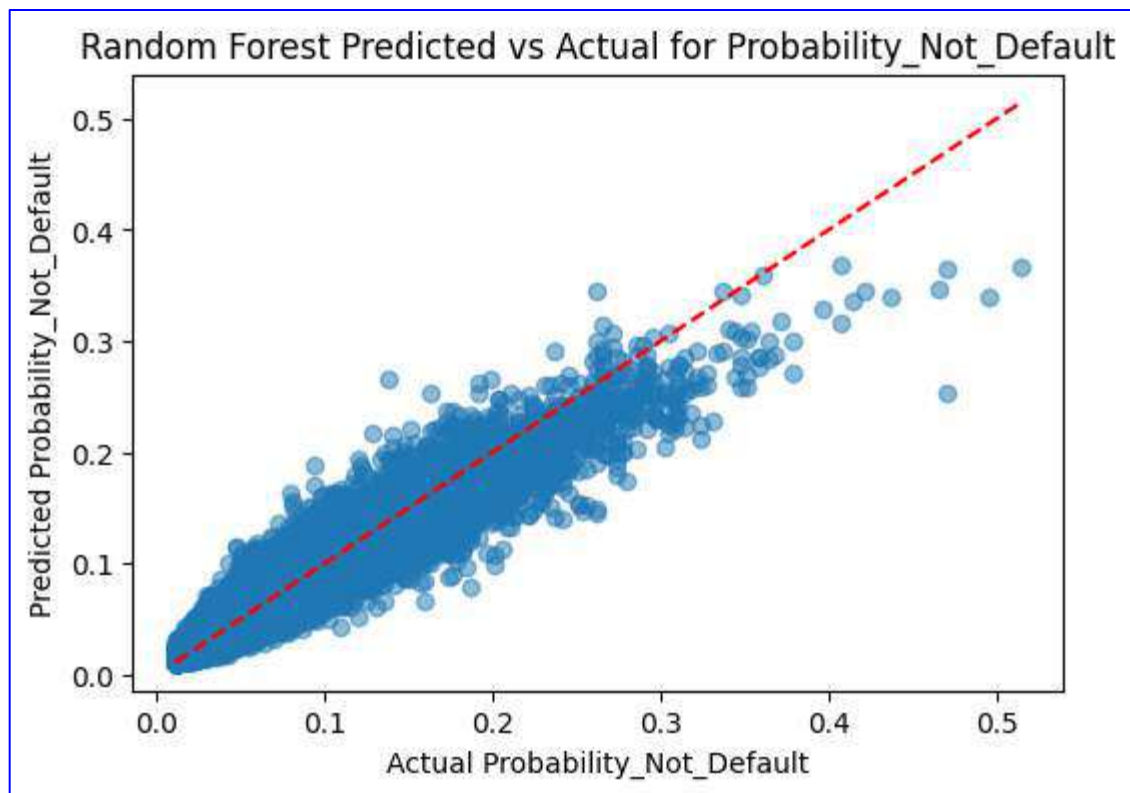


Figure 8.15

❖ **XGBoost**

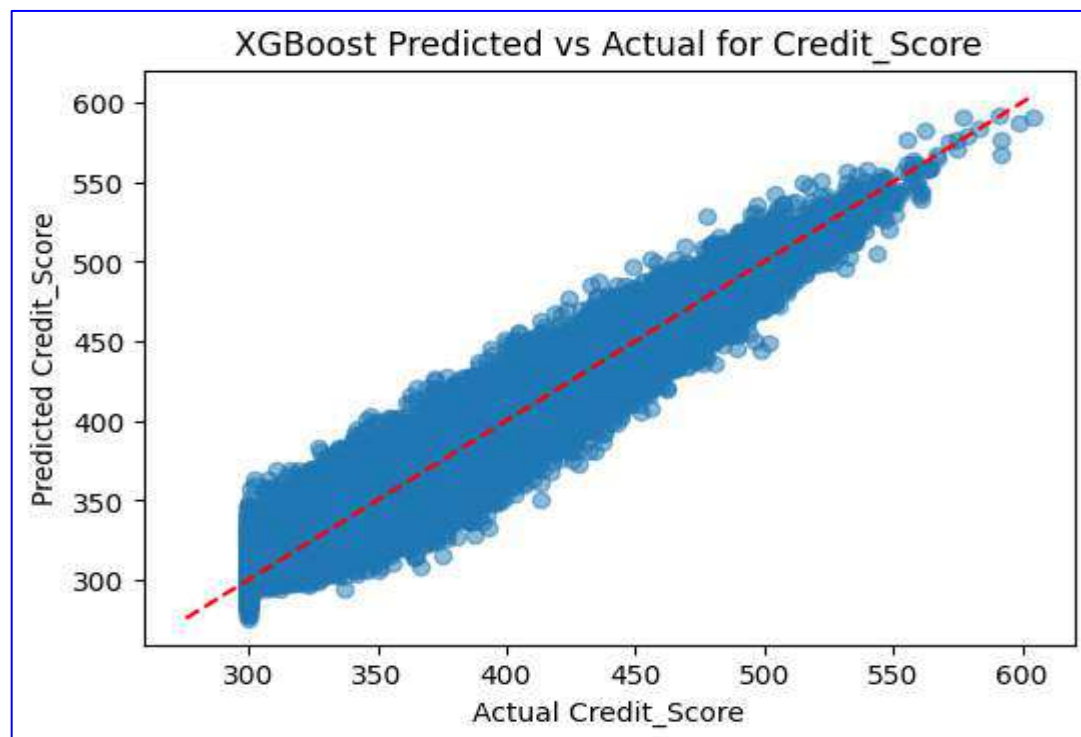


Figure 8.16

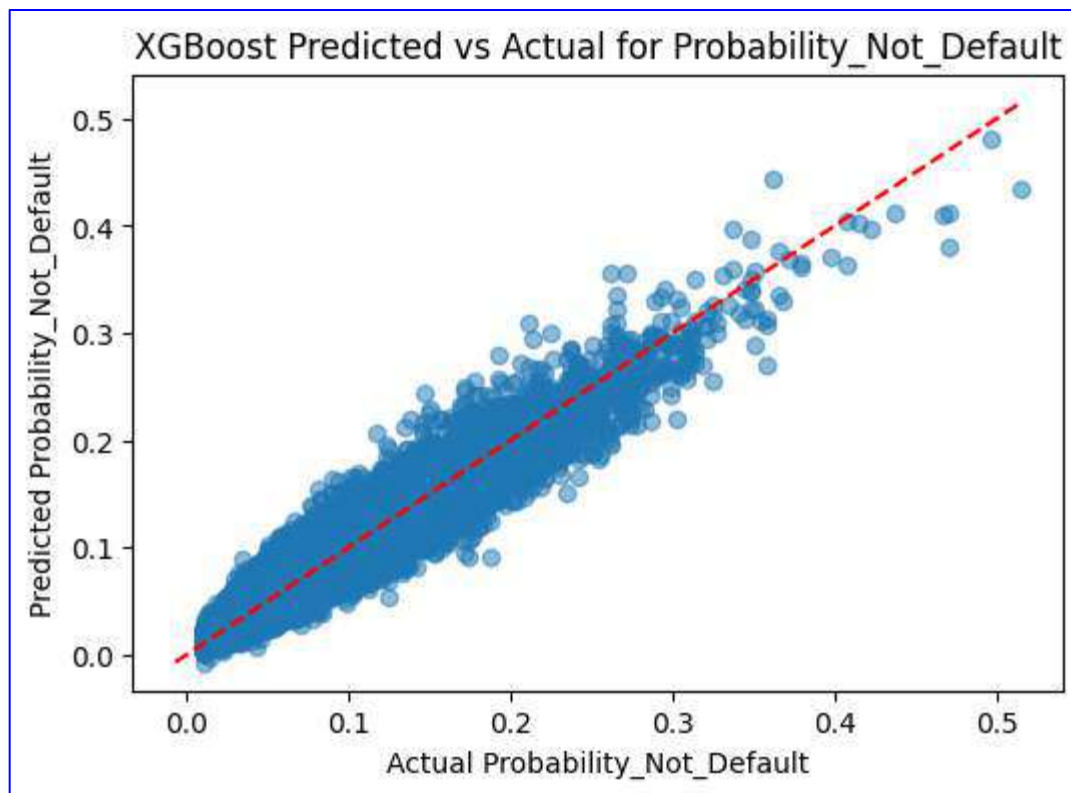


Figure 8.17

❖ **LightGBM**

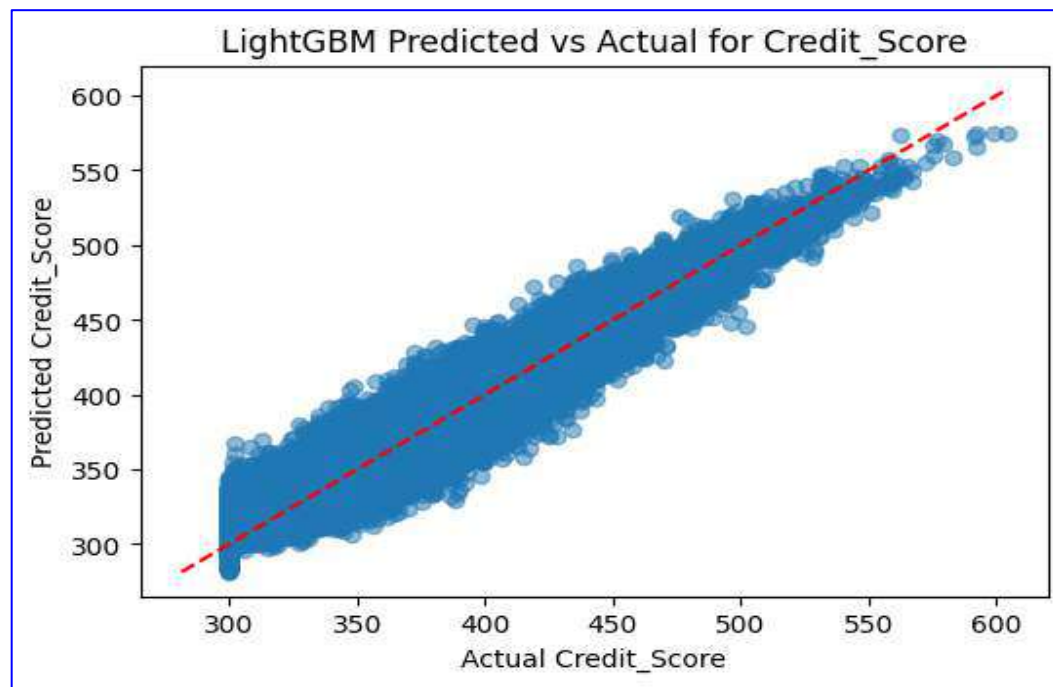


Figure 8.18

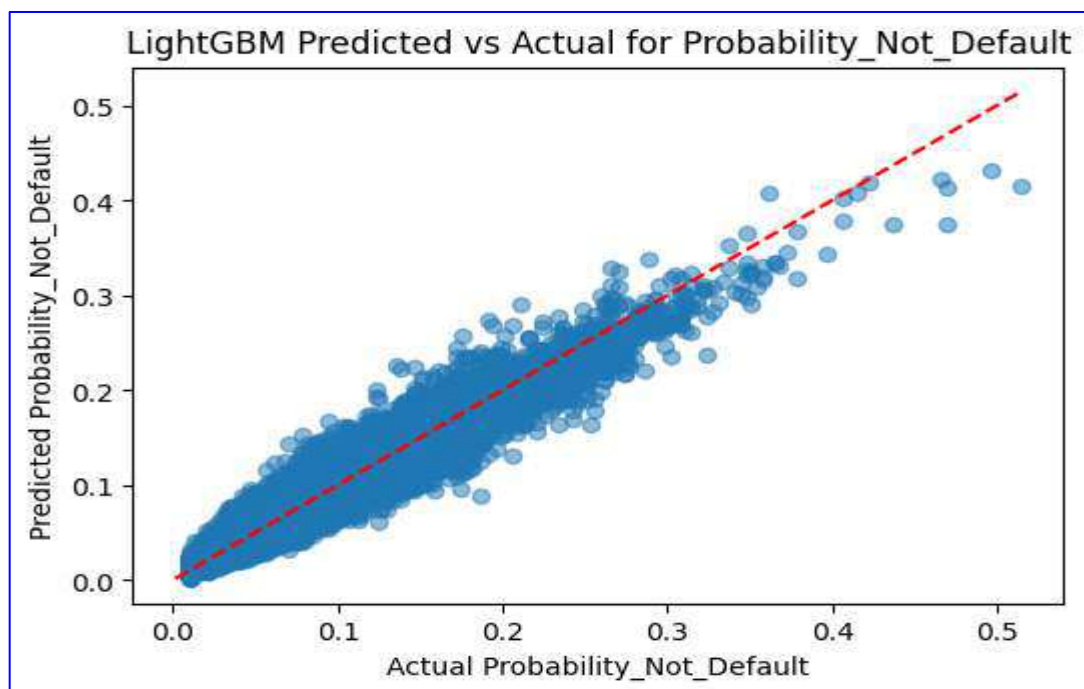


Figure 8.19

❖ MLP(Neural Network)

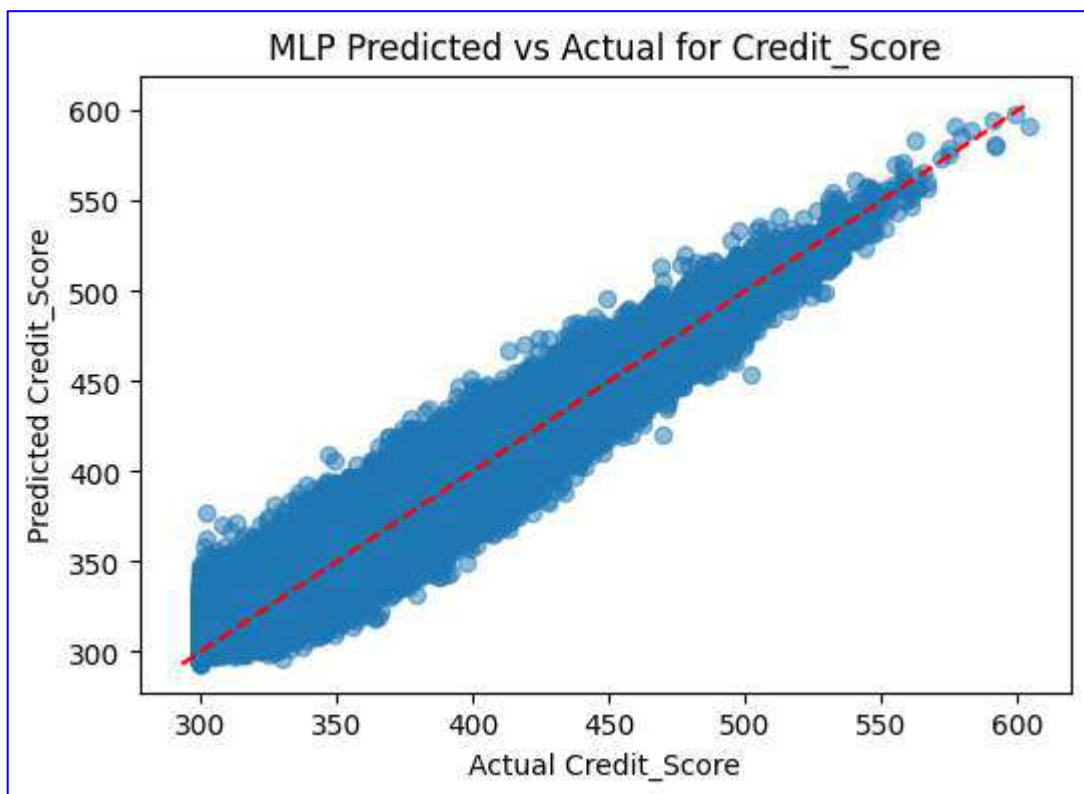


Figure 8.20



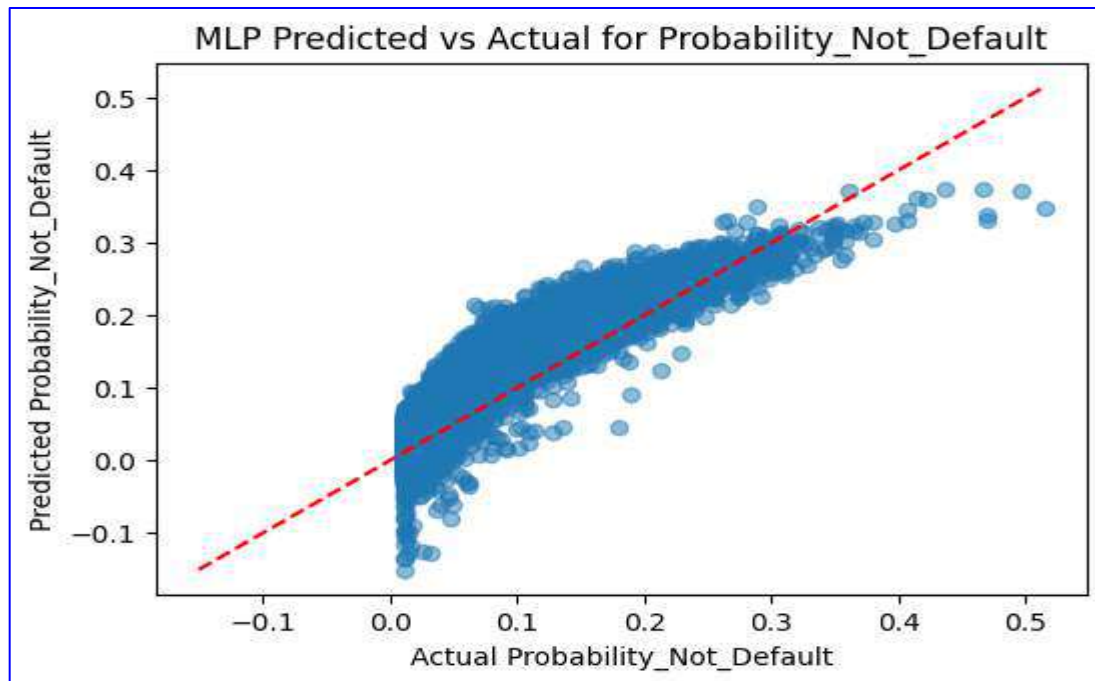


Figure 8.21

❖ 1D CNN

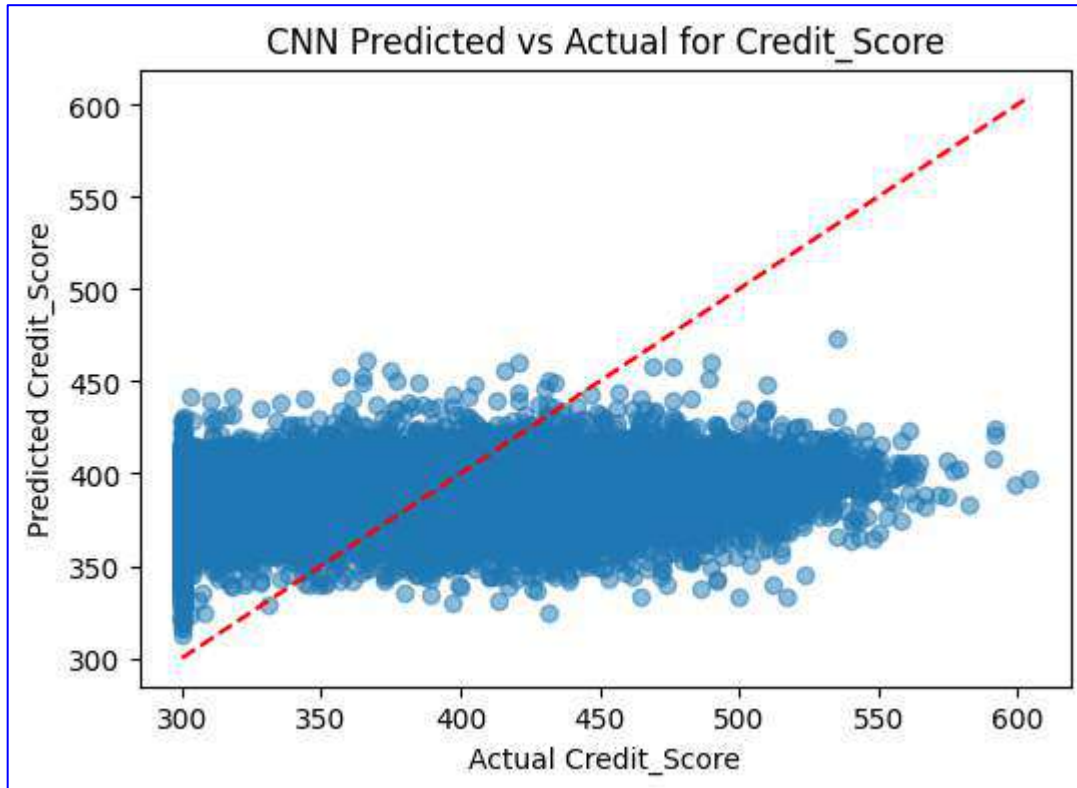


Figure 8.22

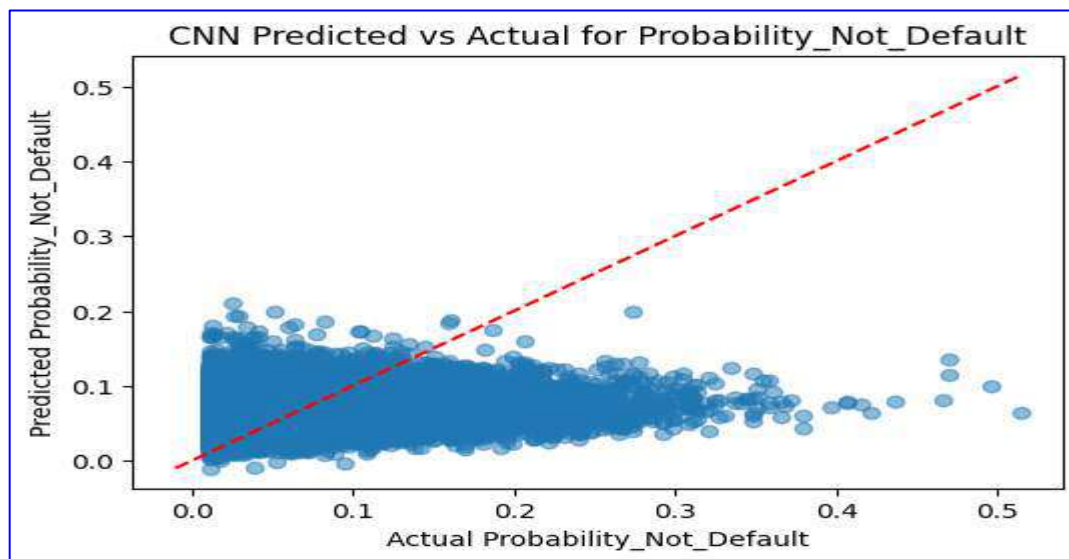


Figure 8.23

❖ LSTM

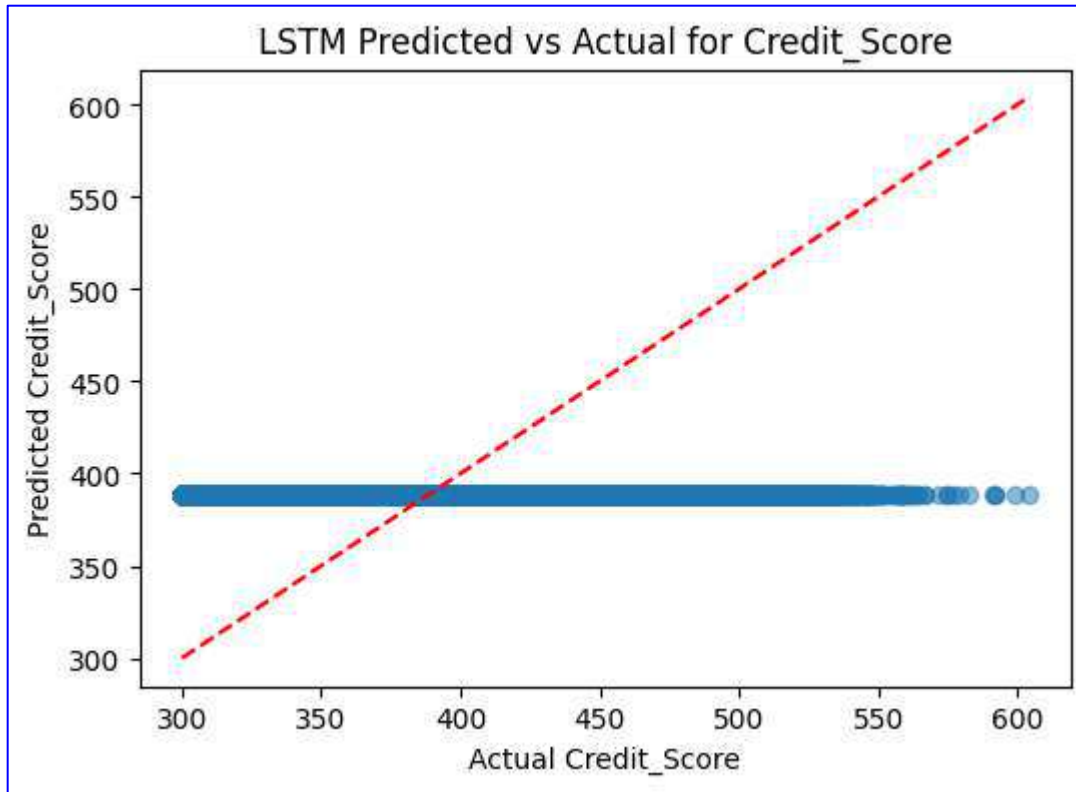


Figure 8.24



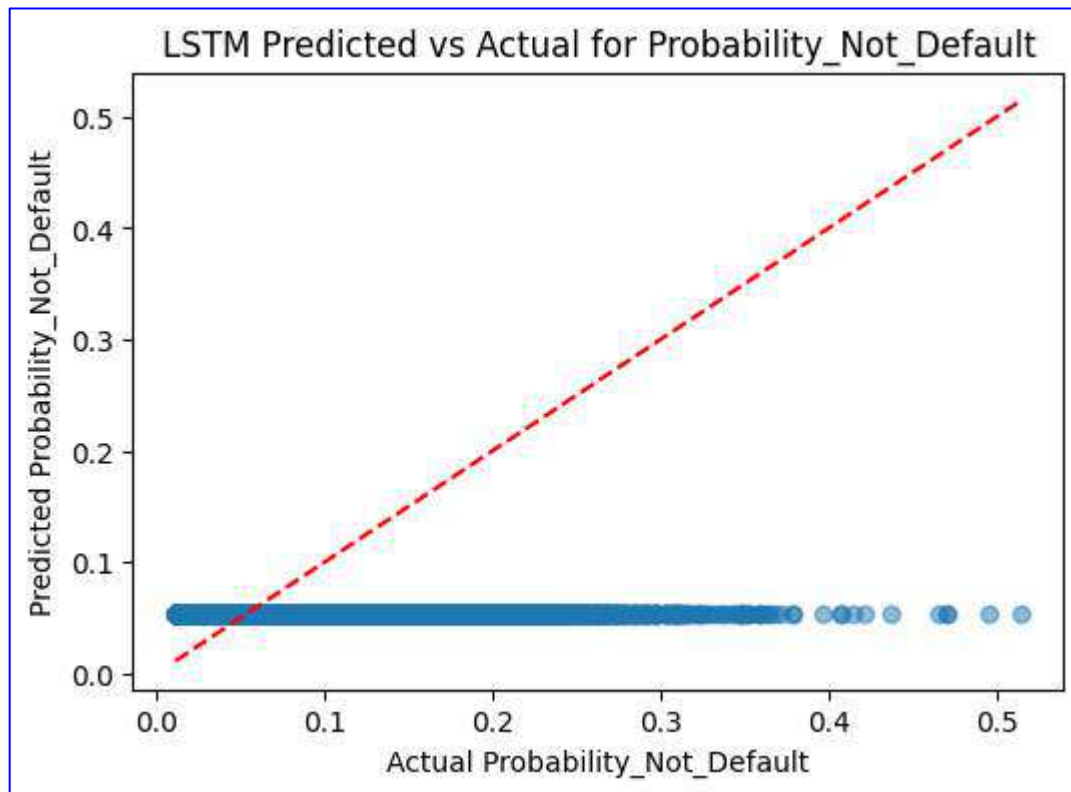


Figure 8.25

So here in each plot actual vs predicted value is plotted for each model and we can see that how model is trying to find best fit line between target variable and features. On the basis of this we can say that 1D CNN and LSTM model are unable to find the best fit line as there are underfitted and whereas for linear regression it is overfitted and unable to find any pattern so instead of learning pattern it memorize the whole data over the time. For Random Forest, XGBoost, LightGBM and MLP these models are able to find the best fit line which describes the data and the plot is also shrinking towards that best fit line over the time so we can say that these models are gradually learning the pattern between the data. So we can use this model for further hyper parameter tuning to improve the performance of the model.

---

## 8.4 Cross-Validation Results

We have done the cross validation using  $k = 5$  I.e 5 folds cross validation to check whether the model is underfitted, overfitted or not. So the results of the cross validation are show as below

	Model	MAE (Credit_Score)	MAE (Probability_Not_Default)	RMSE (Credit_Score)	RMSE (Probability_Not_De
0	Linear Regression	12.3513	0.0170	16.6399	0.0245
1	Random Forest	11.2714	0.0084	14.6360	0.0129
2	XGBoost	10.6501	0.0080	13.7658	0.0118
3	LightGBM	10.2747	0.0076	13.2275	0.0111

Figure 8.26

ot_Default)	RMSE (Credit_Score)	RMSE (Probability_Not_Default)	R2 (Credit_Score)	R2 (Probability_Not_Default)
	16.6399	0.0245	0.9177	0.7413
	14.6360	0.0129	0.9363	0.9282
	13.7658	0.0118	0.9437	0.9403
	13.2275	0.0111	0.9480	0.9470

Figure 8.27

After doing 5 fold cv we found that linear regression is overfitted for Probability of Default as  $R^2$  score for that is very less other models are performing well as they have very good  $R^2$  score.

---

## 8.5 Model Comparison Table

The comparison table for each model on the basis of MAE, RMSE and  $R^2$  score is created for further analysis and then to choose the model for hyper parameter tuning. The comparison table is shown below.

Model Performance Comparison							
	Model	MAE (Credit Score)	RMSE (Credit Score)	$R^2$ (Credit Score)	MAE (Prob Not Default)	RMSE (Prob Not Default)	$R^2$ (Prob Not Default)
0	Linear Regression	12.350718	16.607862	0.919009	0.017065	0.024563	0.742280
1	Random Forest	11.168091	14.548619	0.937848	0.008300	0.012961	0.928247
2	XGBoost	10.588094	13.696138	0.944918	0.007981	0.011819	0.940330
3	LIGHTGBM	10.184439	13.130149	0.949377	10.184439	0.011046	0.947881
4	MLP	9.725981	12.748713	0.952275	0.034037	0.039042	0.348917
5	CNN	47.460945	57.080139	0.043287	0.042085	0.052718	-0.187127
6	LSTM	48.576912	58.362755	-0.000191	0.035385	0.048415	-0.001243

Figure 8.28

So on the basis of this table we can clearly say that CNN and LSTM model are not performing good as they are having very high MAE and RMSE and negative  $R^2$  score values. So this model will not be used for further parameter tuning.

---

## Conclusion

This project sought to build an end-to-end, data-driven credit risk measurement framework that could estimate a quantitative credit score, as well as the probability that a person will not default, to enable more fair and accurate lending. From a heterogeneous set of 45 well-engineered features in the demographic, financial, behavioral, and transactional categories we built a semi-realistic, large dataset simulating the real-world relationships involved in credit risk analysis.

A methodical step-by-step modeling methodology was employed:

- ❖ **Baseline Development** – Linear Regression yielded an interpretable baseline, showing the limitations of strictly linear relationships in being able to capture sophisticated borrower behaviors.
- ❖ **Tree-based ensemble machine learning methods** such as Random Forest, XGBoost, and LightGBM proved to have good predictive strength by successfully extracting intricate non-linear feature relationships.
- ❖ **Fully connected MLP networks** were among the best performing GBMs in this setup, with the potential to be optimized further using more raw, sequential or high-frequency behavioral data. CNN and LSTM models, although poorer in this static tabular setup, are still promising for time-series enriched data.
- ❖ **Rigorous Assessment** – The use of MAE, RMSE, and  $R^2$ , and residual analysis, predicted vs. actual plots, and cross-validation validated a well-balanced and credible model selection process.

### Key Findings:

1. LightGBM proved to be the best overall performer, having the best RMSE and highest  $R^2$  for both targets as well as offering computational efficiency suitable for large-scale use.
2. MLP also performed equally well under certain conditions, indicating the possibility of neural approaches in credit rating if non-linear relations predominate.

---

# Future Scopes

Adding alternative behavior and digital footprint characteristics increased predictive power, allowing for more robust estimates of under-credited or formerly "thin-file" individuals.

## **Practical Applications:**

The model structure that we obtain can be the basis of a successful, transparent, and privacy-friendly credit scoring system. By incorporating that into lending operations, lenders are in a position to:

- While providing credit to a larger population group with lower default risk.
- Make more balanced choices by including non-traditional, but appropriate, behavioral markers.
- Adapt to new information sources and changing borrower behavior.

## **Future Considerations**

Inspired by the excellent results of the proposed modeling techniques, many possibilities can be pursued to improve accuracy, transparency, and use:

### **1. Data Growth and Enrichment**

- ❖ Longitudinal Histories – By including multi-year credit bureau records, payment histories and banking transactions, long-term trends and seasonality of borrower behavior can be acknowledged.
- ❖ Behavioural Data Streams - Involving smartphone usage patterns, location/ mobility scores, and subscription churn rates could provide additional non-traditional forms of credit data.
- ❖ Open Banking and API Integrations - Real-time financial account access via open banking protocols, along with live transactional and balance data.
- ❖ Macroeconomic Indicators - The unemployment rate in the-goal region, inflation or CPI trends, interest rate moves, to contextualize borrower risk profiles.

### **2. Explainability & Compliance**

- ❖ Model-Agnostic Explainability - Global and individual explainability by looking at SHAP, LIME, and counterfactual explainability tools/models.

---

❖ Regulatory Conformity - Compliance with rules in the US Fair Credit Reporting Act (FCRA) and GDPR, as relating to "right to explanation".

❖ Bias detection and mitigation - Fairness measures and bias mitigation algorithm(s) to reduce disparate impact across demographic groups.

### **3. Hybrid Modeling Methods**

❖ GBM + Deep Neural Networks (DNN) - A range of advanced models can be built using gradient boosting for pre-selection and a deep learning model to learn high-order interactions.

❖ Stacked Ensembles- The goal of stacking multiple families of models (LightGBM, MLP, and CatBoost) is again to get the best prediction through the use of meta-learning.

❖ Temporal-Aware Models - Applying all of your sequence models (Temporal CNN's, Transformers) will provide all of your future time-series financial data.

### **4. Deployment & Monitoring**

❖ MLOps Pipeline - Automating data collection, training, validating, deploying and monitoring data through a CI/CD is very efficient.

❖ Real-Time Scoring API - Providing lending companies with API/endpoint access to score lender's deals immediately while doing the application request.

❖ Performance Drift Detection - Detect changes in feature distributions and measure error metric differences that would trigger retraining models before the patients impact their decision quality.

❖ Live A/B Test – Making gradual changes to the new models with comparisons to current live production models to provide incremental uplifts before rolling it out.

### **5. Advanced Feature Engineering**

❖ Network & Graph Features - Map out who the borrower is related to (e.g., co-signer and/or share address) and applying graph analytics will pick up hidden risk communities.

❖ Latent Behavioral Embeddings - Using unsupervised learning (e.g., auto-encoder) will lessen complex behavioral data into a high-density information linking vector.

❖ Risk Segmentation Models - Clustering borrowers into micro-segments and train a model for each for higher knowledge transfer.

---

## REFERENCES

- Jonnalagadda, A. K., & Ramesh Babu, S. (2025). Enhancing credit scoring with alternative data and machine learning for financial inclusion. *South Eastern European Journal of Public Health (SEEJPH)*, XXVI. ISSN: 2197-5248. Posted: 04 January 2025.
- Alamsyah, A., Hafidh, A. A., & Mulya, A. D. (2025). Innovative credit risk assessment: Leveraging social media data for inclusive credit scoring in Indonesia's fintech sector. *Journal of Risk and Financial Management*, 18(2), 74.  
<https://doi.org/10.3390/jrfm18020074>
- Niu, B., Ren, J., & Li, X. (2019). Credit scoring using machine learning by combining social network information: Evidence from peer-to-peer lending. *Information*, 10(12), 397. <https://doi.org/10.3390/info10120397>.
- Bokade, I., Gade, V., Wagh, J., Ingale, S., & Hirve, S. (2025). From credit invisible to credit visible: Interpretable scoring models with alternative data in India. *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, 5(3). <https://doi.org/10.48175/IJARSCT-26385>.