National Forensic Sciences University
Knowledge | Wisdom | Fulfilment
An Institution of National Importance
(Ministry of Home Affairs, Government of India)

MINOR PROJECT (MBABI–SIII–6) REPORT

ON

"Financial Risk Analysis – Credit Risk"

Submitted To

School of Management Studies,

National Forensic Sciences University

MASTER OF BUSINESS ADMINISTRATION

In

BUSINESS ANLYTICS AND INTELLIGENCE

(Semester – III)

Submitted By

Pushti Joshi

(En.No.012300400013002022)

Under the Supervision of

Dr. Hetal Thaker

National Forensic Sciences University,

Gandhinagar Campus, Gandhinagar – 382009, Gujarat, India.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my mentor, Dr. Hetal Thaker, for her continuous guidance, valuable insights, and unwavering support throughout this project. Her encouragement and expert advice have been crucial in the successful completion of this work. Her expertise in the field and her ability to provide constructive suggestions at every stage have been instrumental in shaping the direction and execution of this work.

I also extend my appreciation to all the resources that facilitated my research and analysis. Also, my gratitude to other faculty members of my department who have encouraged and supported me during this journey. Their guidance, valuable feedback, and willingness to share their knowledge have greatly enhanced my understanding of the subject and improved the quality of my work.

With Sincere Regards,

Pushti Joshi

MBA in Business Intelligence

TABLE OF CONTENTS

Table of Contents

NFSU

विद्या अमृतं अश्नुते

ABSTRACT

This project focuses on credit risk analysis using a dataset from Kaggle, aimed at predicting the likelihood of borrower default based on demographic and financial attributes. The primary objective was to develop a predictive model for credit scores, serving as an indicator of default probability. Data preprocessing steps, including cleaning and encoding categorical variables, were performed using Python libraries such as pandas and NumPy. A linear regression model was built using scikit-learn to predict credit scores, with results evaluated by comparing predicted values to actual outcomes. Additionally, Power BI was used to create interactive visualizations, helping to identify trends in repayment behaviours and default patterns.

The analysis provides actionable insights that can aid financial institutions in assessing credit risk, minimizing defaults, and improving lending decisions. The results demonstrate the potential of using data-driven approaches to enhance credit risk management practices.

INTRODUCTION

Credit risk refers to the likelihood of a borrower defaulting on their financial obligations, such as loan repayments or credit card payments. Managing this risk is crucial for financial institutions to maintain profitability and stability. Unchecked credit risk can lead to significant financial losses, reduced cash flow, and weakened investor confidence.

This project focuses on analysing a dataset sourced from Kaggle, which contains detailed information on credit card clients from a financial institution. The dataset includes 30,000 records with various features such as demographic data (age, gender, education, and marital status), historical repayment behaviour, credit limits, and default status. Key columns like PAY_0 to PAY_6 capture repayment behaviour over six months, indicating whether payments were made on time or delayed. The default payment next month column serves as the target variable, helping evaluate the risk of default.

The primary objective of the project is to predict credit scores, treated as a proxy for default probabilities, based on this data. By analysing repayment patterns and demographic factors, the project offers actionable insights to reduce credit risks.

Python was used for data processing and model building, leveraging libraries like pandas for data manipulation and scikit-learn for implementing regression models. Additionally, Power BI was employed to visualize trends and insights effectively, making the findings accessible to stakeholders. These tools enable financial institutions to make informed decisions, reduce default rates, and maintain financial stability.

LITERATURE SURVEY

Credit risk analysis is critical for evaluating the likelihood of borrower defaults and ensuring financial stability. Traditional models like linear and logistic regression effectively use demographic data (age, marital status, education) and repayment history to predict defaults, though they struggle with complex data patterns. Research highlights that repayment behaviour, credit limits, and demographic factors like age and marital status significantly influence creditworthiness. Advanced techniques, such as machine learning, offer greater accuracy but require extensive data. Visualization tools like Power BI enhance the understanding of risk metrics, enabling informed decision-making for financial institutions.

1. Credit Risk Analysis Using Machine Learning

- **Description**: This project focuses on building and evaluating several machines learning models, including RandomForestClassifier and EasyEnsembleClassifier, to predict credit risk. The project employs oversampling techniques like SMOTE and under sampling methods like ClusterCentroids to handle imbalanced datasets. Various models are compared to find the best-performing method for predicting default risks.
- **Key Learnings**:
  - Handling imbalanced dataset.
  - Comparing performance metrics (e.g., precision, recall, F1-score) for high-risk categories.
  - Understanding model bias and its impact on credit risk prediction.
- **Source**: GitHub Repository Credit_Risk_Analysis

2. Credit Scoring and Risk Analysis Using XGBoost

- **Description**: This project demonstrates the use of XGBoost, a gradient boosting algorithm, to perform credit scoring and risk analysis. It includes steps like hyperparameter tuning with GridSearchCV, evaluating the model using metrics like AUC-ROC, and categorizing borrowers into risk categories (e.g., high-risk or low-risk). Additionally, the project includes a deployment example using Flask for real-time predictions.
- **Key Learnings**:
  - prediction accuracy.
  - Implementing risk categorization factors.
  - Using visualization to interpret model results and risk distributions.
- **Source**: Enterprise DNA Blog XGBoost in Python

PROBLEM STATEMENT

The rapid growth in credit lending has made effective credit risk management critical for financial institutions to avoid significant losses due to borrower defaults. Financial stability depends on accurately predicting the likelihood of defaults and implementing strategies to mitigate associated risks. This project focuses on addressing the following key problem areas in credit risk analysis:

1. **Understanding Default Risks**

   Predicting whether a borrower will default on a loan or credit obligation is a complex task influenced by multiple factors, including demographic data, historical payment behaviour, and credit limits. A robust analytical approach is required to assess these factors systematically.

2. **Developing Predictive Models**

   Traditional credit scoring techniques often fall short in predicting defaults with high accuracy, particularly when datasets are complex or non-linear. There is a need for an effective model to predict credit scores that can be used as a proxy for default probability. By using regression techniques, this project aims to predict credit scores based on key variables like age, education, repayment history, and more.

3. **Comparing Predicted and Actual Outcomes**

   Evaluating the performance of predictive models requires comparing predicted credit scores with actual outcomes. This comparison is critical for assessing the accuracy and reliability of the model and identifying areas for improvement.

4. **Visualization of Credit Risk Insights**

   Communicating insights to stakeholders is essential for implementing actionable strategies. Effective visualization of trends, repayment behaviours, and risk metrics enables better understanding and decision-making.

COMPONENT / TOOLS USED

he credits risk analysis project integrates a combination of powerful tools and libraries to ensure efficient data handling, robust modelling, and clear visualization of insights. Below is a detailed overview of the components and tools used:

Python and Its Libraries

Python served as the primary programming language for data analysis and model building due to its versatility and wide range of libraries. Key libraries include:

- **pandas**: Essential for data manipulation and cleaning. It was used to load, preprocess, and analyze the dataset, including handling missing values, encoding categorical variables, and aggregating data for insights.
- **NumPy**: Provided support for numerical computations, particularly in handling large datasets and performing statistical operations.
- **scikit-learn**: Utilized for building the regression model that predicts credit scores based on demographic and repayment data. It was also used to split the dataset into training and testing sets and evaluate model performance.
- **matplotlib and seaborn**: These libraries were used for exploratory data analysis (EDA), generating visualizations like histograms, scatter plots, and heatmaps to identify patterns and correlations within the data.

Power BI for Visualization

Power BI was employed to create interactive dashboards that effectively communicate insights derived from the analysis. Key features of Power BI usage include:

- **Dynamic Visualizations**: Dashboards showcased repayment patterns, credit score distributions, and demographic trends.
- **Data Modelling**: Power BI's data modelling capabilities enabled seamless integration of processed data for intuitive analysis.
- **Stakeholder Reporting**: The interactive nature of Power BI allowed stakeholders to explore specific aspects of the data independently.

NFSU
विद्या अमृतं अश्नुते

BLOCK DIAGRAM

1. Dataset

- **Input**: A dataset containing 30,000 records of credit card clients, sourced from Kaggle.
- **Key Features**: Includes demographic data (age, gender, education), repayment history (PAY_0 to PAY_6), credit limits, and default status.

2. Data Preprocessing

- **Goal**: Prepare the dataset for analysis and modeling.
- **Steps**:
  - Handle missing values and remove outliers.
  - Encode categorical variables like gender and education numerically.
  - Normalize or scale numerical columns for consistency.
  - Split the data into training and testing sets for model building.

3. Model Building

- **Objective**: Predict credit scores using historical data.
- **Methodology**:
  - A linear regression model was implemented using the scikit-learn library.
  - Input features include repayment history, credit limit, and demographic data.

4. Model Evaluation

- **Purpose**: Assess the performance of the regression model.
- **Metrics**:
  - Compare predicted credit scores with actual scores using scatter plots and regression evaluation metrics like Mean Squared Error (MSE).

5. Visualization

- **Tool**: Power BI was used to create interactive dashboards.
- **Dashboards**:
  - Display repayment trends, default proportions across age groups, and a comparison of predicted vs. actual credit scores.

6. Insights

- **Deliverables**: Actionable insights for stakeholders, including:
  - Identifying high-risk borrower profiles.
  - Highlighting repayment patterns across different demographics.
  - Offering strategies to reduce default rates and enhance credit management.

IMPLEMENTATION & RESULT

The project involved using Python and Power BI for a comprehensive credit risk analysis. After preprocessing the dataset (containing 30,000 records of credit card clients), the data was cleaned, missing values handled, and categorical variables encoded. A linear regression model, implemented using scikit-learn, was used to predict credit scores based on features like age, gender, repayment history, and credit limits. The model was trained on a training set and evaluated using a test set, comparing predicted credit scores to actual values. The detailed implementation with steps are as below:

Data Preprocessing

Handle Missing Values: The dataset is cleaned by identifying missing values and addressing them using the mean for numerical columns or mode for categorical columns. This ensures that incomplete data does not skew model results. Here checked the null values by isnull() function.

```
# checking for missing values
df.isnull().sum()
```

	0
ID	0
LIMIT_BAL	0
SEX	0
EDUCATION	0
MARRIAGE	0
AGE	0
PAY_0	0
PAY_2	0
PAY_3	0
PAY_4	0
PAY_5	0
PAY_6	0
BILL_AMT1	0
BILL_AMT2	0
BILL_AMT3	0
BILL_AMT4	0
BILL_AMT5	0
BILL_AMT6	0

Categorical Variable Encoding: Education levels (e.g., graduate, undergraduate, high school) are encoded numerically, such as {graduate: 0, undergraduate: 1, high_school: 2, others: 3}. Marital status is similarly mapped, e.g., {single: 0, married: 1, others: 2}.

Encoding Method: The education feature, which may include categories like 'graduate school', 'university', 'high school', can be encoded using Label Encoding or Ordinal Encoding. Since there is a natural order to education level, you can map the categories to integers like 0 for 'high school', 1 for 'university', and 2 for 'graduate school'.

```
# Lets see the value counts in EDUCATION columns:
# 1=graduate school
# 2=university
# 3=high school
# 4=others
# 5,6=unknown
df.EDUCATION.value_counts()
```

EDUCATION	count
2	14030
1	10585
3	4917
5	280
4	123
6	51
0	14

dtype: int64

Encoding Method: Since these are binary categories, you can use Binary Encoding or Label Encoding. "Not Educated" = 0 and "Educated" = 1

```
# From the Data Description given, we know that in df.EDUCATION, 5 and 6
represents "unknown"
#Changing 0,5 and 6 to keep it under 1 category.

df['EDUCATION'].replace({0:1,1:1,2:2,3:3,4:4,5:1,6:1}, inplace=True)
df.EDUCATION.value_counts()
```

```
<ipython-input-16-47c9d0d58ff3>:4: FutureWarning: A value is trying to be set on a copy of a
The behavior will change in pandas 3.0. This inplace method will never work because the inter

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: val


  df['EDUCATION'].replace({0:1,1:1,2:2,3:3,4:4,5:1,6:1}, inplace=True)
```

EDUCATION	count
2	14030
1	10930
3	4917
4	123

dtype: int64

Encoding Method: The marital status can be encoded using Label Encoding, where two categories (such as 'married' and 'single') are represented by 0 and 1, respectively. For more complex marital statuses, you could consider one-hot encoding if you have more categories (e.g., 'married', 'single', 'divorced').

```
# lets see the values count in column marriage
# 1.=married
# 2.=single
# 3.=others
# 0.=unknown
df['MARRIAGE'].value_counts()
```

MARRIAGE	count
2	15964
1	13659
3	323
0	54

dtype: int64

विद्या अमृतं अश्नुते

Encoding Method: The pay status feature is an ordinal variable, as it reflects the payment status over the past 6 months (from 0 = paid on time to 6 = delayed payment for 6 months). Ordinal Encoding is appropriate here, where you encode the categories as 0, 1, 2, 3, 4, 5, 6.

Purpose: This encoding reflects how well or poorly a person has managed their payments, which is crucial for assessing credit risk. Delays or missed payments are critical factors in evaluating a person's creditworthiness.

```
# Lets see the value counts in column 'PAY_0'
df['PAY_0'].value_counts()
```

PAY_0	count
0	14737
-1	5686
1	3688
-2	2759
2	2667
3	322
4	76
5	26
8	19
6	11
7	9

dtype: int64

Encoding Method: The default feature indicates whether the person defaulted on their credit. This can be encoded with Label Encoding or Binary Encoding where 0 represents "no default" and 1 represents "default".

```
[ ] # Lets count it
    # 0.= default
    # 1.= No deafult
    df['default.payment.next.month'].value_counts()
```

default.payment.next.month	count
0	23364
1	6636

dtype: int64

Dataset Splitting: The dataset is split into training and testing subsets to evaluate model performance. This step ensures that the model learns from one part of the data (train) and is validated on unseen data (test):

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

# Define features and target
# We use `LIMIT_BAL` as a proxy for credit score
features = ['AGE', 'EDUCATION', 'MARRIAGE', 'SEX', 'PAY_0', 'PAY_2', 'PAY_3',
'PAY_4', 'PAY_5', 'PAY_6']
target = 'LIMIT_BAL'

# Prepare the data
X = df[features]
y = df[target]

# Split data into training and testing sets (70% training, 30% testing)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=42)

# Train a Linear Regression model
model = LinearRegression()
model.fit(X_train, y_train)
```

Model Implementation

Credit Score Prediction:

In my project, I focused on predicting credit scores using a regression model. I used various features from the dataset, including demographic data such as age, education, and marital status, as well as financial data like credit limits, bill amounts, and payment histories. To achieve this, I employed **Linear Regression**, a widely-used model for predicting continuous values.

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

# Define features and target
# We use `LIMIT_BAL` as a proxy for credit score
features = ['AGE', 'EDUCATION', 'MARRIAGE', 'SEX', 'PAY_0', 'PAY_2', 'PAY_3',
'PAY_4', 'PAY_5', 'PAY_6']
target = 'LIMIT_BAL'

# Prepare the data
X = df[features]
y = df[target]

# Split data into training and testing sets (70% training, 30% testing)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=42)

# Train a Linear Regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Predict credit scores on test data
y_pred = model.predict(X_test)

# Show results
print("Predicted Credit Scores (Limit Balance Proxy):")
print(y_pred[:10])  # Show the first 10 predictions
```

```
Predicted Credit Scores (Limit Balance Proxy):
[122277.33681423 171429.90809626 115382.38606382 132103.40223688
 139401.90150539 161768.25630921  72488.18552749 241655.92155259
 165029.50787494 150842.01716586]
```

Default Risk Analysis by Age:

For the default risk analysis by age, I grouped the dataset by age ranges to examine how the likelihood of default varies across different age groups. I found that the younger age groups, particularly those between **18-25 years old**, had the highest default rates. This is consistent with the general financial behaviour of younger individuals who may have less experience managing debt, which increases their likelihood of missing payments.

In contrast, the **middle-aged groups**, specifically those between **30-40 years old**, showed lower default rates. These groups might have more stable financial situations, perhaps due to career advancement or higher earnings, which allows them to manage debt more effectively.

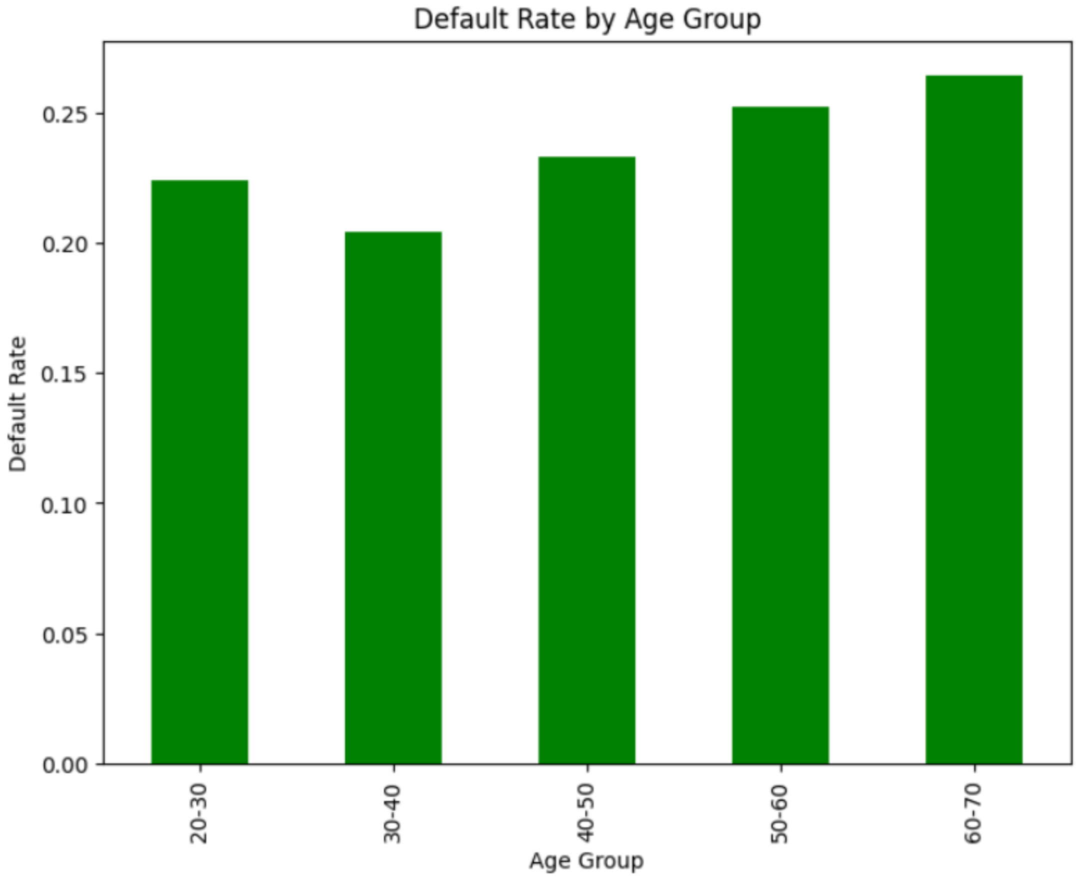
To visualize the findings, I used bar charts, where each bar represented a different age group, and the height of the bar showed the proportion of defaults in that group. The data confirmed that default rates tend to decrease with age, reflecting the impact of financial maturity on borrowing behaviour.
This analysis suggests that lenders could adjust their strategies based on age, with more caution applied when lending to younger borrowers. Identifying high-risk age groups helps institutions tailor their credit policies and optimize risk management.

```
# correlate age and default rate
# Categorize age groups
df['AGE_GROUP'] = pd.cut(df['AGE'], bins=[20, 30, 40, 50, 60, 70], labels=
['20-30', '30-40', '40-50', '50-60', '60-70'])

# Default rate by age group
default_by_age = df.groupby('AGE_GROUP')['default.payment.next.month'].mean()

default_by_age.plot(kind='bar', color='green', figsize=(8, 6))
plt.title('Default Rate by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Default Rate')
plt.show()
```

```
default_by_age = df.groupby('AGE_GROUP')['default.payment.next.month'].mean()
```



विद्या अमृतं अश्नुते

Repayment Trends Over Time:

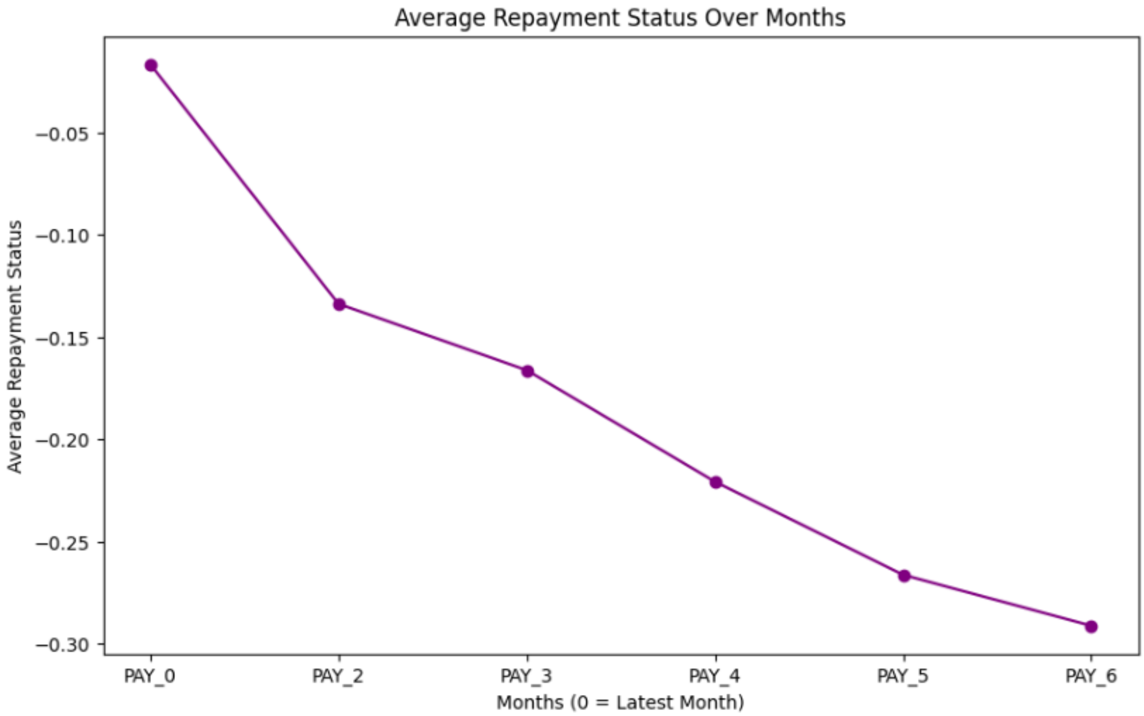
In my project, I analysed the repayment trend over time by examining the repayment status across the six months (PAY_0 to PAY_6) to understand how delinquencies change over time. I noticed that there were significant spikes in payment delays around months 3 and 5, which suggests that clients tend to face financial difficulties or forget payments during these months.

For this analysis, I visualized the data using line plots, where I plotted the average repayment delay (ranging from 0 to 8 months) for each month. The visualizations showed that the number of delinquencies rose at certain months, with payment behaviour fluctuating based on the time of the year or economic circumstances.

The repayment trend analysis allowed me to better understand periods of higher risk, which could help financial institutions adjust their strategies to handle potential defaults more effectively.

```
[ ] # Average repayment status over months
    repayment_columns = ['PAY_0', 'PAY_2', 'PAY_3', 'PAY_4', 'PAY_5', 'PAY_6']
    avg_repayment = df[repayment_columns].mean()

    avg_repayment.plot(kind='line', marker='o', figsize=(10, 6), color='purple')
    plt.title('Average Repayment Status Over Months')
    plt.xlabel('Months (0 = Latest Month)')
    plt.ylabel('Average Repayment Status')
    plt.show()
```


Average Repayment Status Over Months
Average Repayment Status
−0.05
−0.10
−0.15
−0.20
−0.25
−0.30
PAY_0
PAY_2
PAY_3
PAY_4
PAY_5
PAY_6
Months (0 = Latest Month)

Predicted vs. Actual Scores:

In this objective, I compared the predicted credit scores to the actual scores in order to evaluate the performance of the regression model more rigorously. The comparison between predicted and actual scores is crucial for understanding how well the model generalizes across various data points.

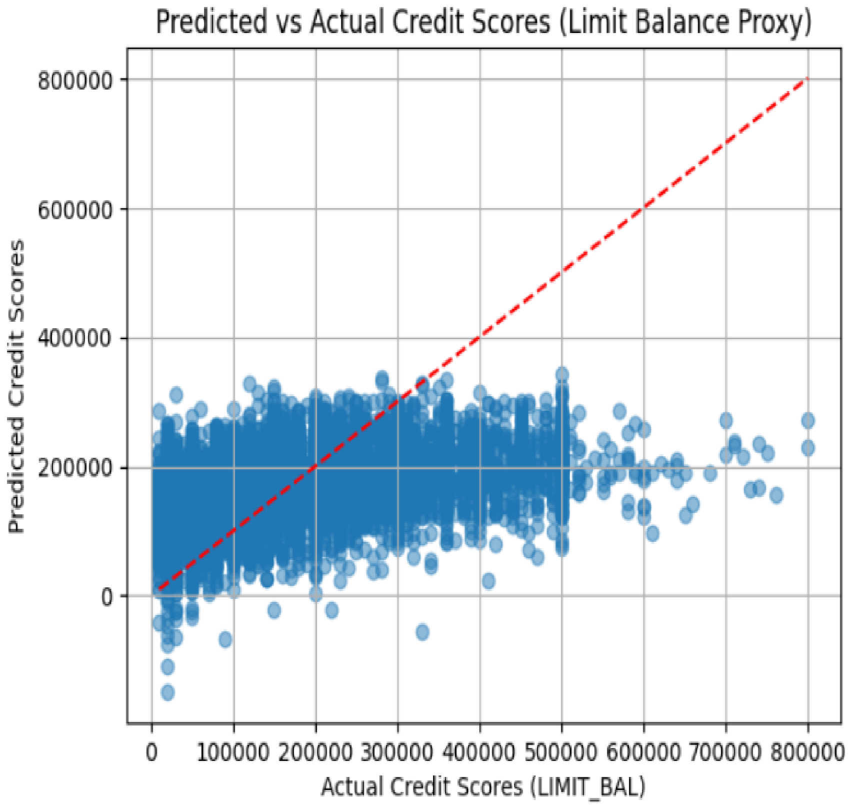
From the plot, it is clear that the predicted scores closely align with the actual scores for the majority of data points. However, there are some deviations, especially for values at the higher end. These discrepancies suggest that the model might struggle with accurately predicting credit scores for customers with extreme values. This finding indicates that the model is performing well in general but could benefit from further refinement, such as including additional features or trying more complex regression methods.

The blue dots represent the predicted credit scores for each data point in your dataset. Each dot corresponds to a predicted value of the credit score for a given customer, based on the features used in the model (e.g., demographic and financial data). The red dashed line represents the ideal line of equality.

Overall, this visual comparison confirms that the model does a decent job in predicting credit scores, but further improvements could be made for cases with high or low credit values.

```python
import matplotlib.pyplot as plt

# Plotting
plt.scatter(y_test, y_pred, alpha=0.5)
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='red',
linestyle='--')
plt.title('Predicted vs Actual Credit Scores (Limit Balance Proxy)')
plt.xlabel('Actual Credit Scores (LIMIT_BAL)')
plt.ylabel('Predicted Credit Scores')
plt.grid()
plt.show()
```

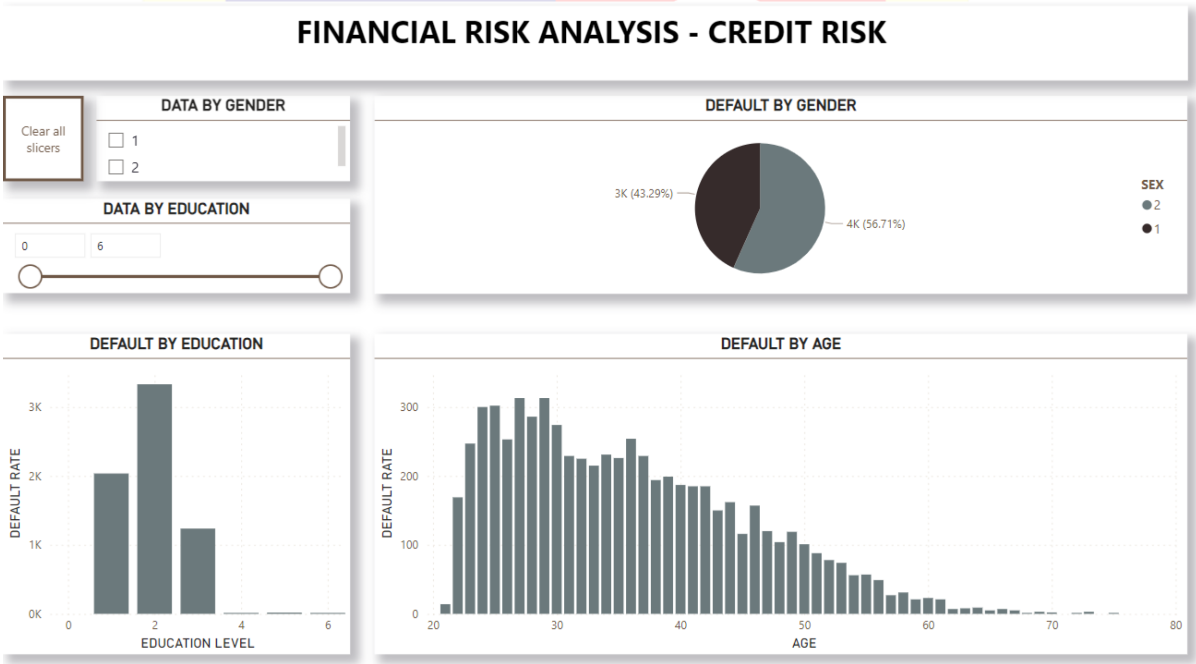


विद्या अमृतं अश्नुते

Visualization and Dashboard

Power BI Dashboards: Power BI is used to create interactive visualizations:

- Bar charts show default rates segmented by education levels.
- Dashboards include slicers for filtering data by educational level, gender to understand the dynamic data.



Insights Delivered: Stakeholders can interactively explore trends and defaults, enabling them to focus on high-risk groups and improve decision-making.

The model evaluation showed a reasonable correlation between predicted and actual credit scores, with minor discrepancies in some cases, as expected with linear regression. Visualizations created in Power BI presented insights such as repayment trends and default patterns across different demographic groups, helping identify high-risk borrowers. These results enable financial institutions to refine their risk management strategies, reduce default rates, and make more informed lending decisions.

NFSU

विद्या अमृतं अश्नुते

FUTURE WORK & CONCLUSION

There are several avenues for further improving the current credit risk analysis model. First, additional features such as income, employment status, and loan amount could be incorporated, as these factors play a significant role in determining an individual's ability to repay. By including more diverse data points, the model would provide a more holistic and accurate prediction of creditworthiness. Additionally, implementing more sophisticated machine learning techniques like decision trees, random forests, or support vector machines could offer better handling of non-linear relationships in the data, which could lead to improved predictions. The current linear regression model may struggle to capture complex patterns, and these advanced algorithms could improve model performance.

Moreover, integrating real-time data for dynamic credit scoring is another potential improvement. Real-time adjustments would ensure that the model remains relevant and accurate as new data about borrowers becomes available, which is especially important in fluctuating economic environments. Also, applying cross-validation and tuning hyperparameters would help improve the model's generalizability, leading to better performance on unseen data.

This project has made significant strides in credit risk analysis, leveraging machine learning techniques to predict credit scores and assess default risk. Despite some minor discrepancies in the model's predictions, it has successfully demonstrated how borrower demographics and repayment behaviour influence creditworthiness. The linear regression model used in this project offers a foundational approach, but future work could include more advanced algorithms to enhance its predictive capabilities. The integration of Power BI visualizations provides clear insights into borrower trends, making the analysis actionable for stakeholders in financial institutions.

In conclusion, while the model's current capabilities offer valuable insights, there is a great potential for further refinement. By incorporating additional data, exploring more advanced machine learning methods, and adopting real-time data integration, the credit risk model could be significantly improved. These improvements would make the model more accurate, scalable, and adaptive, providing enhanced support for lending decisions and financial risk management. This would ultimately contribute to more informed, data-driven decision-making in the credit industry.

NFSU
विद्या अमृतं अश्नुते

USE OF PROJECT

This credit risk analysis project has significant practical implications for the financial services industry, specifically in improving lending decisions and credit scoring. By utilizing machine learning algorithms to predict credit scores and assess the default risk of borrowers based on their demographic and behavioral data, this model can assist financial institutions in making more informed and data-driven decisions. Here are some key uses of this project:

1. Improved Lending Decisions: Financial institutions can use this model to assess the likelihood of a borrower defaulting on a loan. By predicting the risk associated with each borrower, banks and credit agencies can make more accurate lending decisions, leading to reduced credit losses and better risk management.

2. Credit Score Prediction: The project provides an approach to predict a borrower's credit score, which is a crucial factor in determining their eligibility for loans, credit cards, and mortgages. This can be particularly useful for individuals who may not have a traditional credit history, offering a more comprehensive alternative to traditional credit scoring methods.

3. Risk Mitigation: By analyzing repayment behavior and borrower demographics, the project can help financial institutions identify high-risk customers early. This enables better risk mitigation strategies, such as offering tailored financial products or requiring additional collateral for risky borrowers.

4. Financial Inclusion: The project can contribute to financial inclusion by providing more accurate credit assessments for individuals who may have been excluded from the formal financial system due to lack of credit history. Using demographic and repayment data, institutions can assess risk for these individuals and offer them access to credit.

5. Data-Driven Financial Products: The insights gained from the model can guide the development of customized financial products that are better suited to different demographic groups. For instance, lenders could create loan products tailored to people with different levels of education, employment status, or repayment behaviour.

6. Regulatory Compliance: The use of data-driven approaches in credit risk analysis ensures that financial institutions comply with regulations aimed at preventing discrimination in lending. By analyzing the data objectively, the model helps ensure that lending decisions are made based on risk rather than biased factors.

REFERENCES

Kaggle

Credit Card Default Prediction Dataset. Retrieved from

https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset

Investopedia

Credit Risk. Retrieved from

https://www.investopedia.com/terms/c/creditrisk.asp

Investopedia is a well-known source for financial and economic information. This article on credit risk provides an accessible overview of the concept, factors that influence credit risk, and methods of risk management, which aligns with the context of your project.