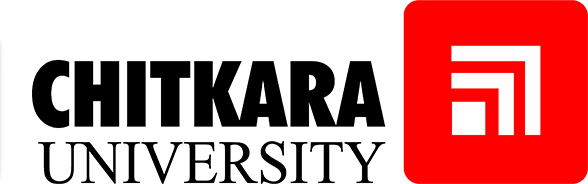
**Capstone Project**

Project Report Semester-IV (Batch-2022)

**Loan Default Risk Prediction**



# Supervised By: Submitted By:

Mr. Mohd. Talib Sir

Yakshit Singla 2210992551 (G-28)

Yash 2210992553 (G-28)

Yash Goyal 2210992554 (G-28)

Yash Goyal 2210992555 (G-28)

**Department of Computer Science and Engineering** **Chitkara University Institute of Engineering & Technology,**

**Chitkara University, Punjab**

**Table Of Contents**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Title** | **Page No.** |
| **1.** | Introduction | 3 |
| **2.** | Problem Statement | 4 |
| **3.** | Software Interaction | 5 |
| **3.** | Exploratory Data Analysis | 6-9 |
| **5.** | Data Preprocessing | 10 |
| **6.** | ML - Models | 11-14 |
| **7.** | Conclusion | 15 |

**Introduction**

In the realm of finance, managing the risk associated with lending is paramount for ensuring the stability and sustainability of financial institutions. With the advent of machine learning techniques, the ability to accurately assess and predict loan default risk has seen significant advancements, revolutionizing the lending landscape.

This project delves into the application of machine learning algorithms to predict loan default risk, leveraging a comprehensive dataset encompassing various borrower attributes, loan features, and historical repayment patterns. By harnessing the power of data-driven insights, this endeavor aims to provide lenders with a robust framework for making informed decisions, mitigating the inherent uncertainties associated with lending operations.

The significance of this project lies in its potential to enhance the efficiency and efficacy of lending practices, enabling financial institutions to proactively identify and manage default risks while simultaneously fostering responsible lending behavior. Through the utilization of advanced analytics and predictive modeling, this project endeavors to contribute to the evolution of credit risk management, paving the way for more resilient and adaptive financial ecosystems.

In the subsequent sections, we will explore the methodology employed, the dataset utilized, the features engineered, the model selection process, and the evaluation metrics employed to gauge the performance of the predictive models. Additionally, we will discuss the implications of the findings and potential avenues for future research, thereby providing a comprehensive overview of the endeavor and its implications for the financial industry.

**Problem Statement**

The objective of this project is to develop a robust machine learning model capable of accurately predicting the likelihood of loan default for individual borrowers. In the context of lending institutions, accurately assessing the risk associated with extending credit is essential for maintaining financial stability and minimizing potential losses. Therefore, the primary aim of this project is to address the following key challenges:

Risk Assessment Precision: Develop a predictive model that can effectively differentiate between borrowers who are likely to default on their loans and those who are not, with a high degree of accuracy. This involves identifying relevant features and leveraging advanced machine learning algorithms to discern subtle patterns indicative of default risk.

Data Integration and Preprocessing: Integrate and preprocess a diverse array of data sources, including borrower demographics, loan attributes, credit history, and economic indicators, to construct a comprehensive dataset suitable for training and testing the predictive model. This entails handling missing values, encoding categorical variables, and conducting feature engineering to enhance predictive performance.

Model Generalization and Interpretability: Build a model that not only exhibits strong predictive performance on the training dataset but also generalizes well to unseen data. Additionally, strive to enhance model interpretability to facilitate stakeholder understanding and trust in the decision-making process, thereby enabling lenders to make informed judgments regarding loan approval and risk management strategies.

**Software interaction:**

- **Jupyter Notebook** serves as the primary Integrated Development Environment (IDE) for its interactive and collaborative features.

- **Pandas** and **NumPy** play integral roles in data manipulation, pre-processing, and mathematical operations, respectively.

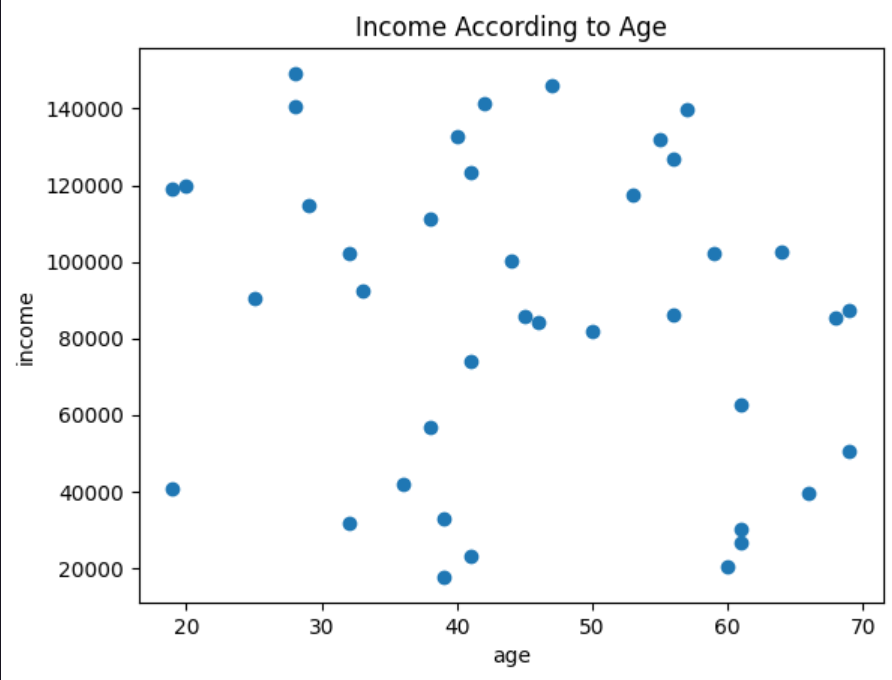
- **Automated Exploratory Data Analysis (EDA**) is facilitated through data prep, streamlining insights discovery.

- **Matplotlib**, **Seaborn,** and **Plotly** are utilized for visualization, enabling the creation of compelling graphical representations.

- **GitHub** serves as the cornerstone for version control, ensuring collaboration, change tracking, and repository integrity.

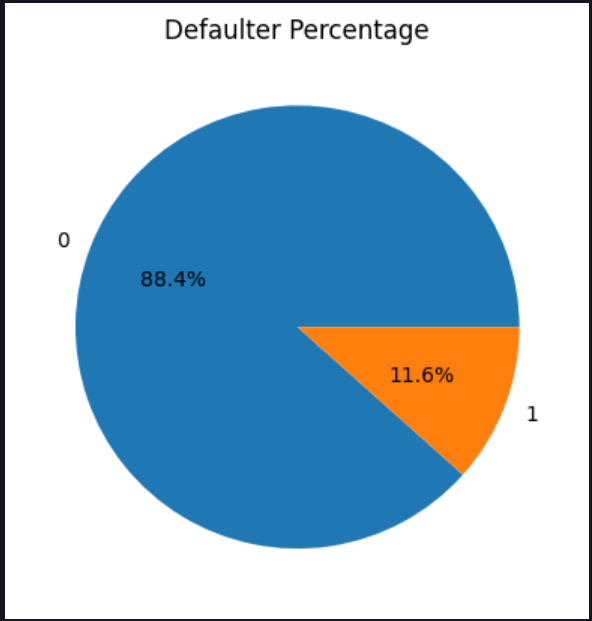
- **Sklearn** Scikit-learn is utilized for implementing machine learning algorithms and statistical models in Python.

**Exploratory Data Analysis**

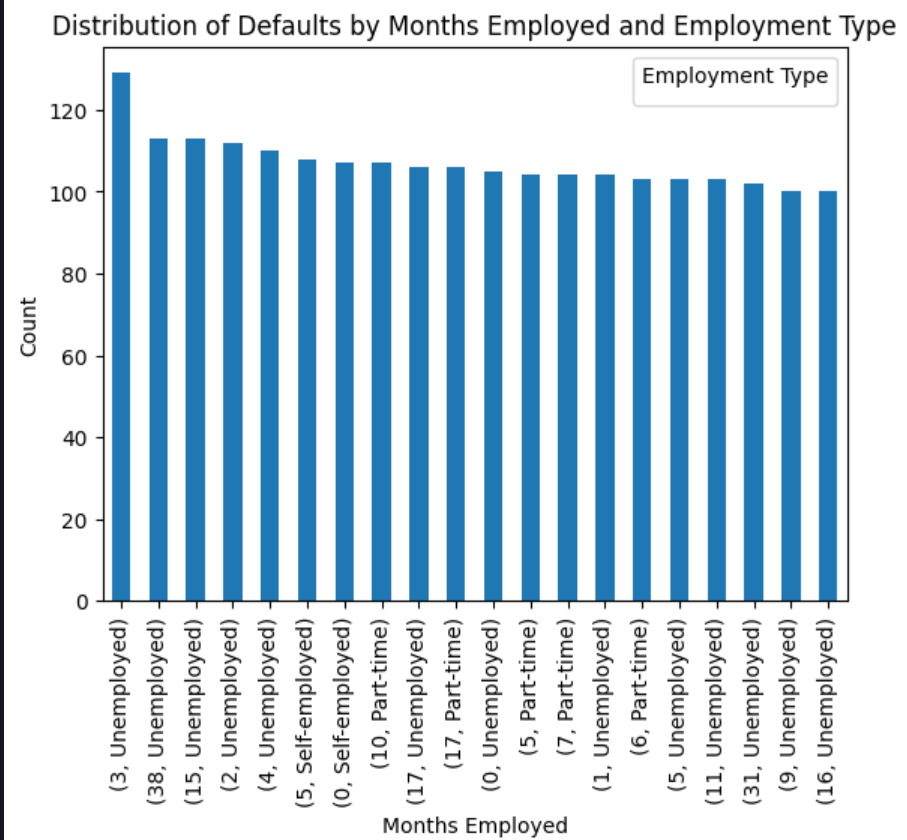


Distribution of Blood Pressure

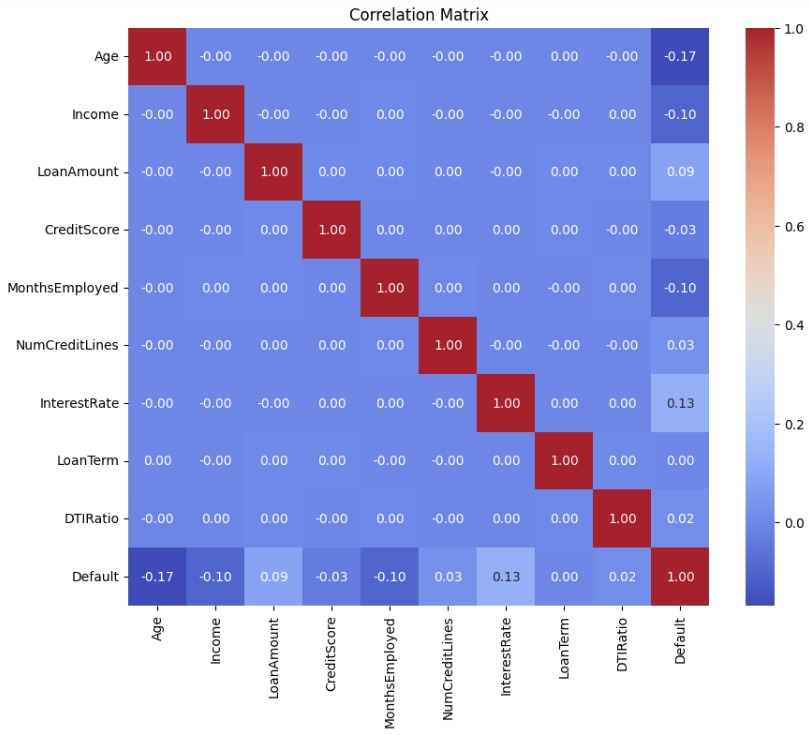
Blood Pressure and Income Distribution



Pie Chart for Defaulter Distribution



Months Employed Distribution



Heatmap of Dataset

**Data Preprocessing**

1. **Handling Missing Values:**

The code drops two columns ("Unnamed: 0" and "id") using drop() function. Then, it transforms the target variable "satisfaction" into binary values. Next, it standardizes column names for consistency. However, there's no explicit handling of missing values in the provided snippet. Additional steps such as imputation or deletion might be needed for robust data preprocessing.

1. **Handling Outliers:**

This function identifies outliers within the DataFrame `df` based on the provided `features`. It calculates quartiles and the Interquartile Range (IQR) for each feature, then identifies outlier indices falling beyond 1.5 times the IQR from the quartiles. Outlier indices occurring more than twice across features are stored and returned as `multiple\_outliers`.

**3. Handling Categorical Values:**

This code transforms categorical columns in `train\_df` into binary indicators, assigning 1 to rows matching the first unique value and 0 otherwise. Then, it converts the "satisfaction" column in `test\_df` to binary, representing "satisfied" as 1 and others as 0.

**ML – MODEL 1**

**Logistic Regression:**

1.Data Preparation:

Import necessary libraries (e.g., NumPy, pandas).

Load or create your dataset.

Split the dataset into features (X) and target variable (y).

2.Import the logistic regression model from a library (e.g., scikit-learn).

Create an instance of the logistic regression model.

3.Fit the model to the training data.

Use the .fit() method of the logistic regression model, passing the features (X\_train) and target variable (y\_train).

4.Use the trained model to make predictions on test data.

Use the .predict() method of the logistic regression model, passing the features (X\_test).

Evaluation:

5.Calculate metrics such as Precision, Recall or others depending on the problem using Confusion Matrix.

**ML – MODEL 2**

**Decision Tree Classifier:**

1. Import the DecisionTreeClassifier from sklearn.tree.

2. Instantiate a DecisionTreeClassifier object as dt.

3. Define a parameter grid containing potential hyperparameters for tuning.

4. Instantiate a GridSearchCV object (clf) with the DecisionTreeClassifier and the defined parameter grid.

5. Fit the GridSearchCV object to the training data (x\_train, y\_train).

6. Print the best parameters found by the grid search.

7. Make predictions (y\_pred) on the test data using the best model found.

8. Evaluate the model's performance using the accuracy\_score function and print the result.

**ML – MODEL 3**

**SUPPORT VECTOR MACHINE(SVM):**

1.Import the svm model from a library (e.g., scikit-learn).

2.Create an instance of the svm model.

3.Fit the model to the training data.

Use the .fit() method of svm model, passing the features (X\_train) and target variable (y\_train).

4.Use the trained model to make predictions on test data.

Use the .predict() method of the svm model, passing the features (X\_test).

Evaluation:

5.Kernels in SVM can be used to transform data into higher-dimensional spaces, allowing for nonlinear separation of classes.

**ML – MODEL 4**

1.Import the KNN model from a library (e.g., scikit-learn).

2.Create an instance of the KNN model.

3.Fit the model to the training data.

Use the .fit() method of KNN model, passing the features (X\_train) and target variable (y\_train).

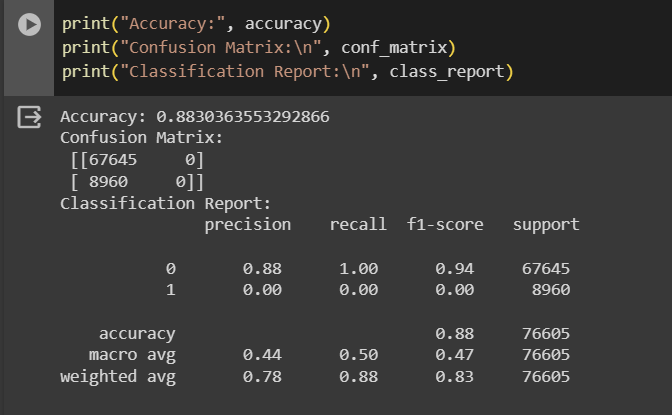
4.Use the trained model to make predictions on test data.

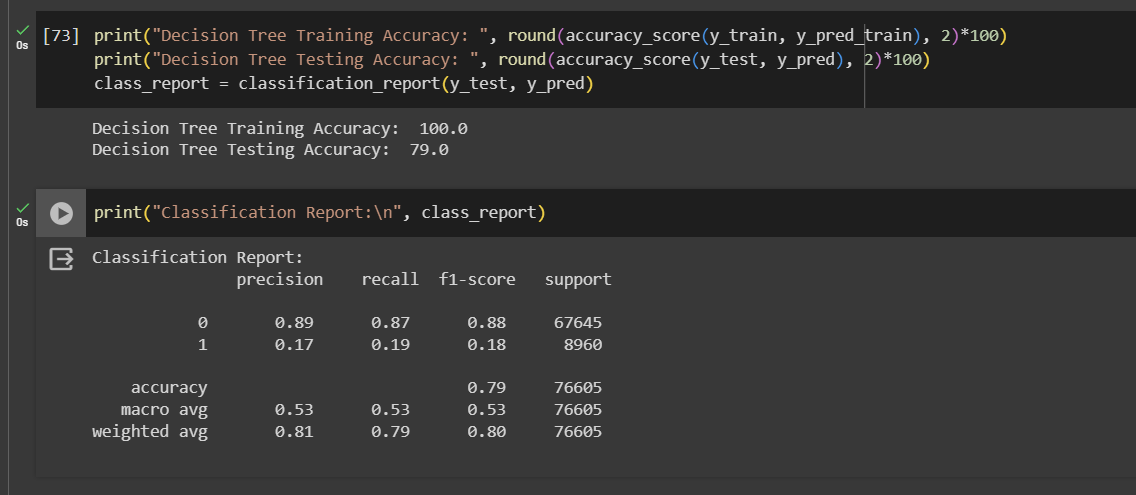
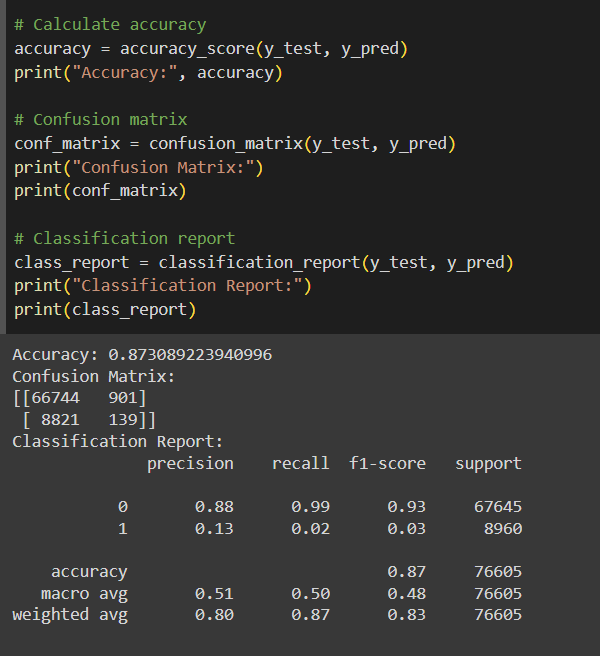
Use the .predict() method of the KNN model, passing the features (X\_test).

Evaluation:

5.Evaluate the model's performance using the accuracy, confusion matrix, classification report function and print the result.

**Accuracy results of each ML Model:-**



**Conclusion**

In this project, we embarked on a journey to develop a machine learning model for predicting loan default risk, with the overarching goal of enhancing the efficiency and effectiveness of lending operations while mitigating potential losses for financial institutions. Through the utilization of advanced analytics and predictive modeling techniques, we endeavored to address the complex challenges inherent in credit risk assessment and management.

Our efforts culminated in the construction of a robust predictive model capable of accurately discerning the likelihood of loan default for individual borrowers. Leveraging a comprehensive dataset encompassing various borrower attributes, loan features, and historical repayment patterns, we employed state-of-the-art machine learning algorithms to extract meaningful insights and identify predictive patterns indicative of default risk.

In conclusion, this project represents a significant step forward in the quest to harness the power of data-driven insights for optimizing lending practices and fostering financial resilience. By leveraging the synergies between machine learning and finance, we embark on a path towards a more inclusive, efficient, and sustainable lending ecosystem.