REPORT ON DEFAULT PREDICTION USING LOAN DATA

DharmpratapSingh Vaghela

1. Introduction

Loan default prediction is a crucial area of research in financial risk management, as it helps financial institutions assess and mitigate the risks associated with lending. The primary objective of this project was to build a machine learning model that predicts whether a loan applicant is likely to default based on their demographic, financial, and loan-related information.

The dataset used in this project was obtained from Kaggle, which provided anonymized data on loan applicants, including details such as age, income, employment length, loan amount, and purpose of the loan. By analyzing this data, we aimed to uncover significant insights about the factors that influence default risk and develop models that could accurately predict defaults.

This report provides a detailed narrative of the project's lifecycle, from data exploration and preprocessing to model building, evaluation, and interpretation. Each step is discussed in depth to offer a clear understanding of the methodology and the rationale behind the decisions made during the project.

2. Dataset Description

The dataset was sourced from Kaggle. https://www.kaggle.com/datasets/ganjerlawrence/loan-risk-prediction-dataset/data

Columns and Descriptions

ID	Unique identifier for each loan applicant.
Age	Age of the loan applicant.
Income	Annual income of the loan applicant.
Home	Home ownership status (Own, Mortgage, Rent).
Emp_Length	Employment length in years.
Intent	Purpose of the loan (e.g., education, home improvement).
Amount	Loan amount applied for.
Rate	Interest rate on the loan.
Status	Loan approval status (Fully Paid, Charged Off, Current).
Percent_Income	Loan amount as a percentage of income.
Default	Whether the applicant has defaulted on a loan previously (Yes/No).
Cred_Length	Length of the applicant's credit history in years.

The dataset contained 8,145 records, with a mix of numerical and categorical variables. The goal was to predict the "Default" status based on the other features.

3. Exploratory Data Analysis (EDA)

Missing Values

- **Emp_Length**: Imputed missing values with the **mode** to maintain consistency in the ordinal nature of the variable.
- **Rate**: Imputed missing values with the **median** to handle outliers and skewed distributions effectively.

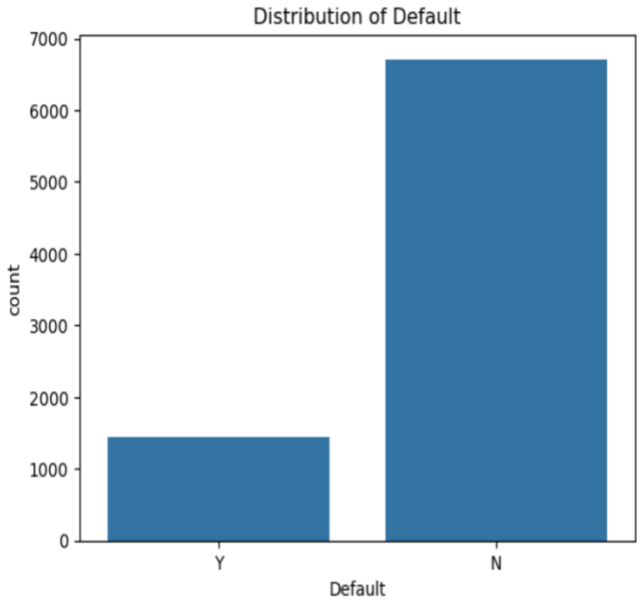
Insights from Features:

1. **Emp_Length**:
   - Ordinal in nature; mode imputation was chosen as it aligns with the distribution.
   - Most applicants have an employment length of 5 years or less.
2. **Rate**:
   - Continuous variable; median imputation was used due to robustness against outliers.
   - The interest rate distribution is slightly skewed, with a typical value around the median.
3. **Default Distribution**:
   - The dataset is imbalanced, with most loans being non-default (Default = No).

3.1 Understanding the Target Variable

The target variable, **Default**, was found to be highly imbalanced, with a majority of applicants labeled as "No Default." Specifically, out of 8,145 records, only 1,435 applicants had a history of default, while 6,710 did not. This imbalance posed a challenge for model training, as most machine learning algorithms tend to favor the majority class.

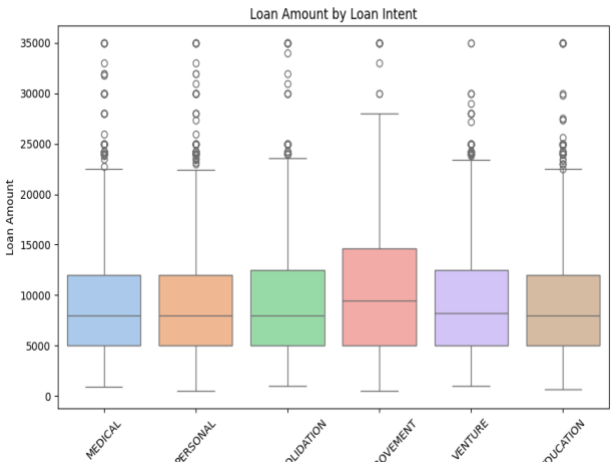
A bar chart was created to visualize this imbalance, which highlighted the need for techniques such as oversampling to address the issue.



Loan Amount by Intent

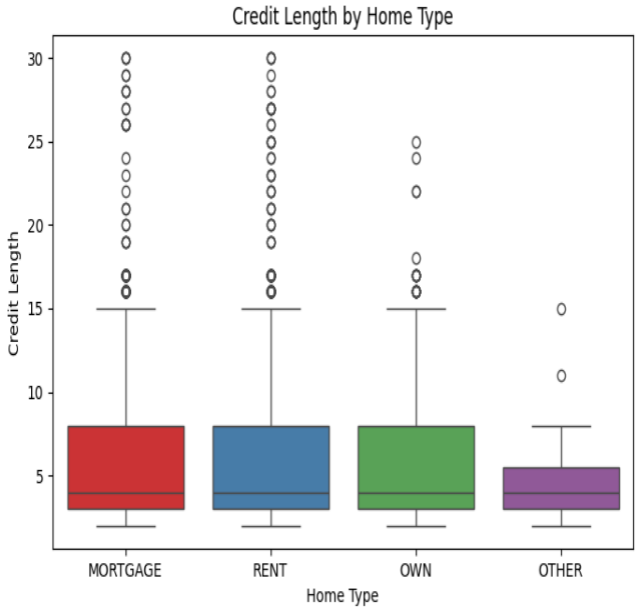
A boxplot was created to explore the distribution of loan amounts across different loan purposes. It revealed:

- Higher median loan amounts for "Debt Consolidation" loans, reflecting the financial nature of these requests.
- Greater variance in loan amounts for "Education" and "Home Improvement" purposes, likely due to varying financial needs associated with these categories.



Credit Length by Home Type

The relationship between credit length and home ownership status was examined using a boxplot. Applicants with mortgages generally had longer credit histories, while renters had shorter histories. This pattern aligns with the notion that individuals with longer credit histories are more likely to own homes.

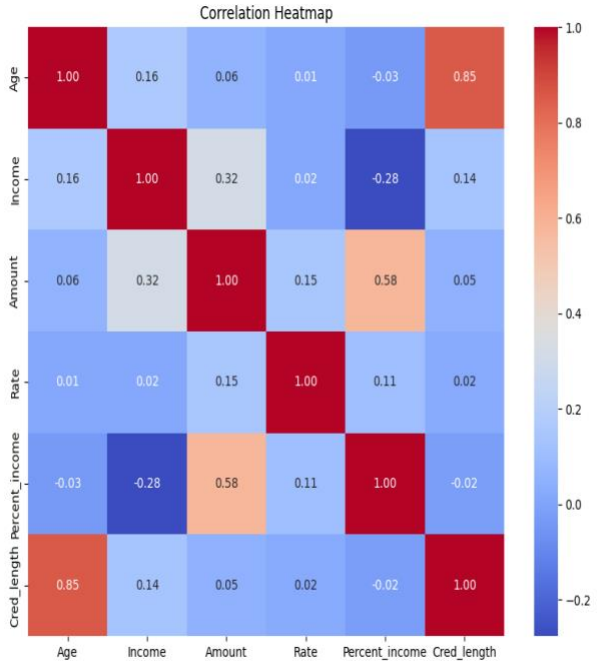


Correlation Heatmap

A correlation heatmap was generated to examine the relationships between numerical variables. Key findings included:

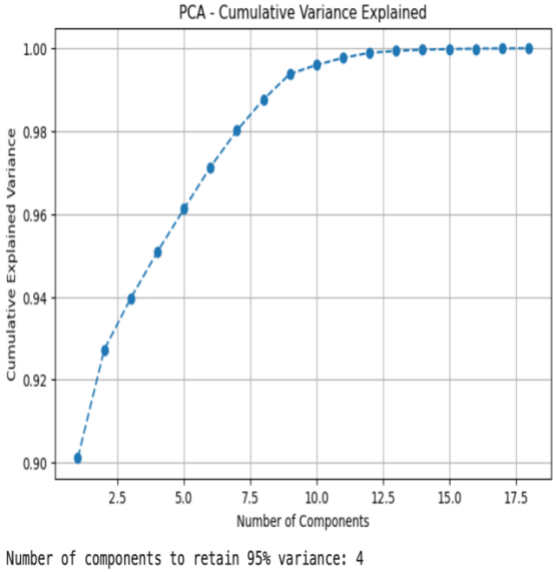
- A strong positive correlation between **Age** and **Cred_Length**, indicating that older applicants generally have longer credit histories.
- Moderate correlations between **Income** and **Amount**, suggesting that higher-income applicants tend to request larger loans.
- Weak correlations between most features and the target variable, implying potential non-linear relationships.

The correlation matrix helped identify potential multicollinearity issues and guided the decision to apply dimensionality reduction techniques like PCA.



3.2 Principal Component Analysis (PCA)

To address potential multicollinearity and reduce the dimensionality of the dataset, Principal Component Analysis (PCA) was applied. By analyzing the explained variance, it was determined that four principal components were sufficient to capture 95% of the variance in the data. This step helped simplify the dataset while retaining its essential information.



Number of components to retain 95% variance: 4

4. Data Preprocessing

Data preprocessing was a critical step in preparing the dataset for model training. It involved the following tasks:

4.1 Handling Missing Values

- The **Emp_Length** column had 236 missing values, which were imputed using the mode, as this column represents an ordinal variable.
- The **Rate** column had 762 missing values, which were imputed using the median to avoid the influence of outliers.

4.2 Encoding Categorical Variables

One-hot encoding was applied to the **Home** and **Intent** columns to convert them into numerical formats. This approach created binary columns for each category, enabling machine learning models to process the data effectively.

4.3 Scaling

Min-Max scaling was used to normalize the numerical columns, ensuring that all features were on a comparable scale. This step was especially important for models like logistic regression and neural networks.

4.4 Addressing Class Imbalance

The dataset's class imbalance was addressed using the Synthetic Minority Oversampling Technique (SMOTE). By oversampling the minority class, we created a balanced dataset that improved the performance of the models.

5. Models and Evaluation

Several machine learning models were trained and evaluated to identify the best-performing algorithm for predicting loan defaults. The models included Logistic Regression, Lasso Regression, Random Forest, XGBoost, Support Vector Machines (SVM), and Neural Networks.

Performance Metrics

The models were evaluated using the following metrics:

- **Accuracy**: The overall correctness of the predictions.
- **Precision**: The proportion of true positives among predicted positives.
- **Recall**: The proportion of true positives among actual positives.

- **F1-Score**: The harmonic mean of precision and recall.
- **ROC-AUC**: A measure of the model's ability to distinguish between classes.

Evaluation Metrics

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.86	0.86	0.86	0.86	0.92
Lasso Regression	0.86	0.87	0.87	0.86	0.91
Random Forest	0.84	0.84	0.84	0.84	0.92
XGBoost	0.81	0.82	0.81	0.81	0.88
Neural Networks	0.60	0.60	0.60	0.60	-
SVM	0.57	0.57	0.57	0.56	0.60

The Random Forest and Logistic Regression models emerged as the top performers, both achieving an ROC-AUC of 0.92. However, Random Forest demonstrated better handling of non-linear relationships, while Logistic Regression provided better interpretability.

6. Hyperparameter Tuning

Random Forest:

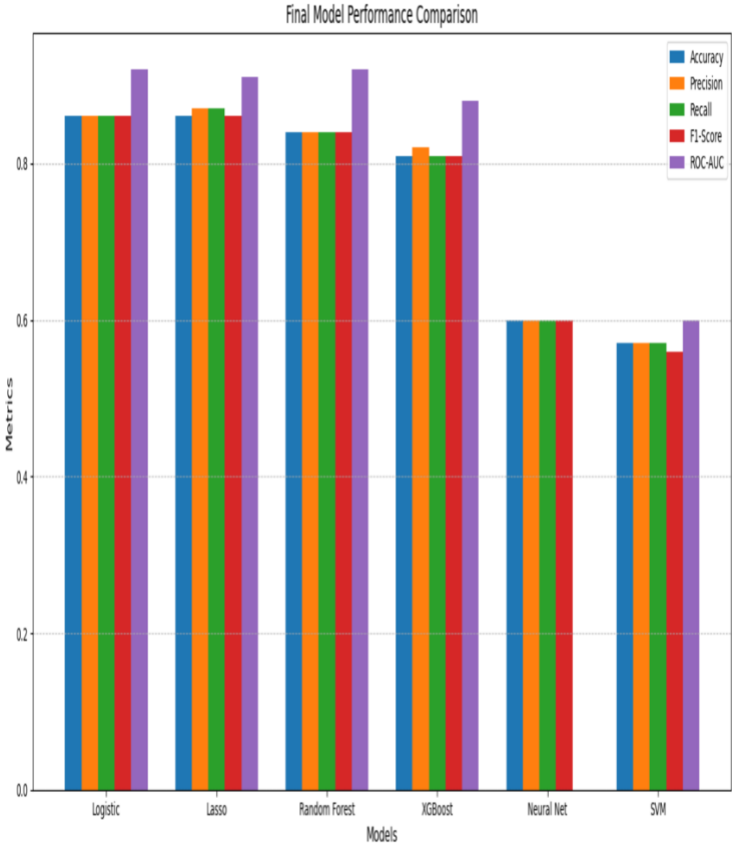
- Tuned n_estimators, max_depth, min_samples_split, and min_samples_leaf using GridSearchCV.
- Best parameters improved ROC-AUC to **0.92**.

XGBoost:

- Tuned learning_rate, max_depth, subsample, and colsample_bytree using RandomizedSearchCV.
- Best parameters improved ROC-AUC to **0.88**.

7. Results and Insights

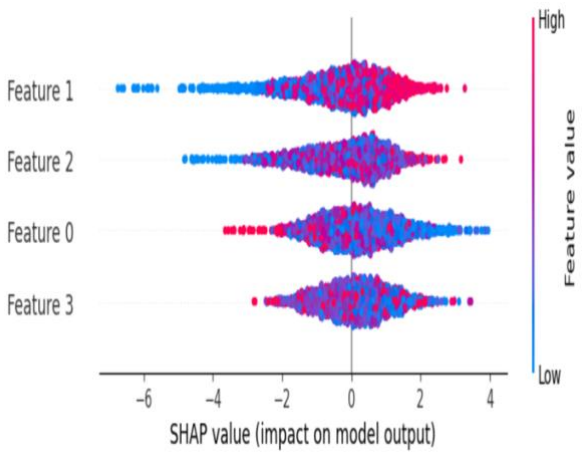
1. **Best Model**:
   - Random Forest and Logistic Regression with dataset_linear performed the best.
   - Both achieved a **ROC-AUC of 0.92**.



2. **Interpretability**:
   - Logistic Regression provided clear insights into the impact of features.
   - Random Forest and XGBoost highlighted the importance of non-linear patterns.
3. **Neural Networks**:
   - Requires further tuning and more data for competitive results.

Feature Importance:

- SHAP values were used for interpretability.



8. Conclusion

This project successfully demonstrated the application of machine learning techniques to predict loan defaults. Key takeaways include:

- The importance of addressing class imbalance in imbalanced datasets.
- The effectiveness of Random Forest in capturing non-linear relationships.
- The value of interpretability offered by Logistic Regression.