LOAN DEFAULT PREDICTION

Saahil Shaikh	21070122132
Sanjana Sharma	21070122142
Sanna Johnson	21070122145
Sarina Khatri	21070122147

Table of Contents

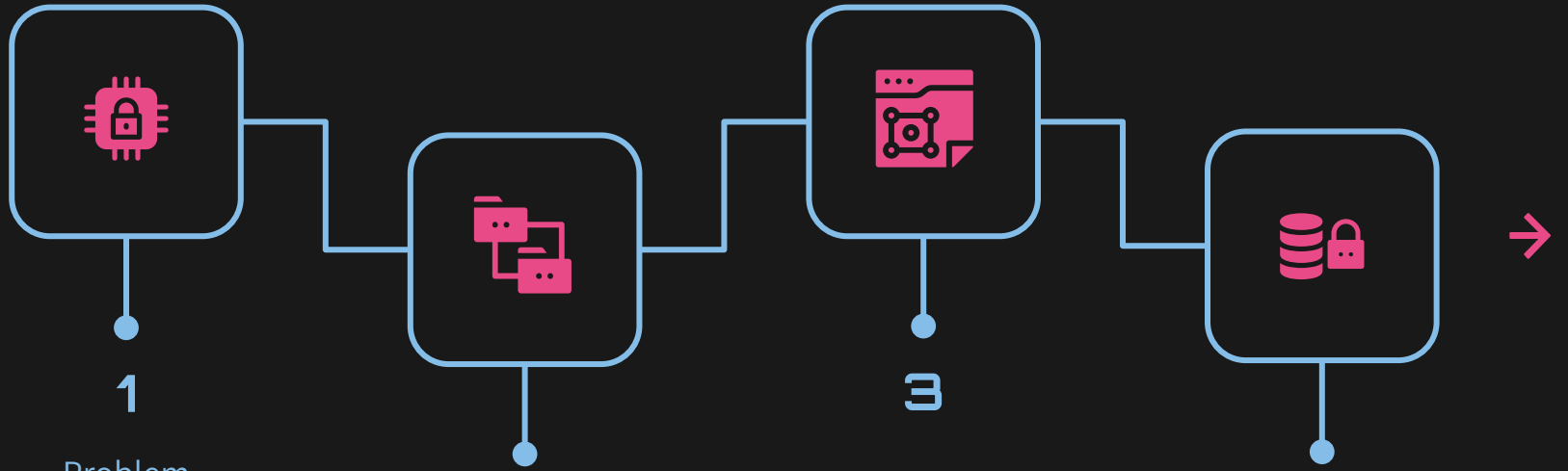
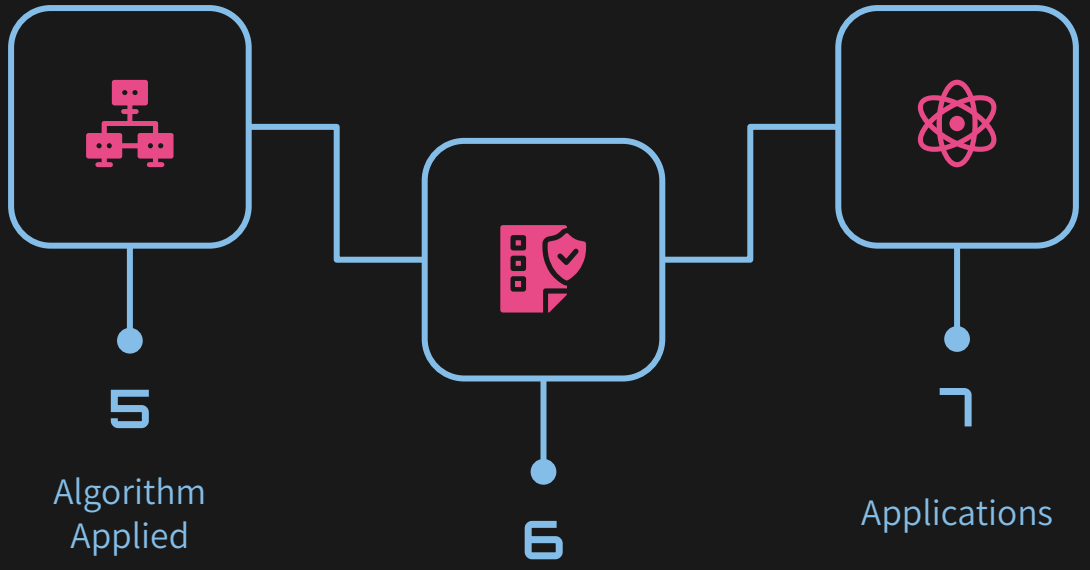


Table of Contents



Problem Statement

Based on an in-depth analysis of a comprehensive dataset containing historical loan applicant information, financial profiles, credit history, and loan repayment behavior, the objective of this data science project is to develop a robust and accurate loan default prediction model. The goal is to provide lending institutions with a reliable tool for assessing and managing the risk associated with loan approvals, thereby enhancing their ability to make informed and responsible lending decisions.

NEED

01

Risk Mitigation

Lending institutions face significant financial risks from loan defaults, and an effective prediction model can mitigate these risks by identifying high-risk borrowers and reducing potential losses.

02

Responsible Lending

This project supports responsible lending by accurately assessing applicants' creditworthiness, promoting fair access to credit while preventing excessive debt.
here

03

Improved Decision Making

A robust prediction model enables data-driven loan approval decisions, improving lending profitability and sustainability.

NEED

04
Regulatory Compliance

A well-constructed prediction model helps lending institutions comply with regulatory risk assessment requirements, avoiding penalties.

05
Customer Experience

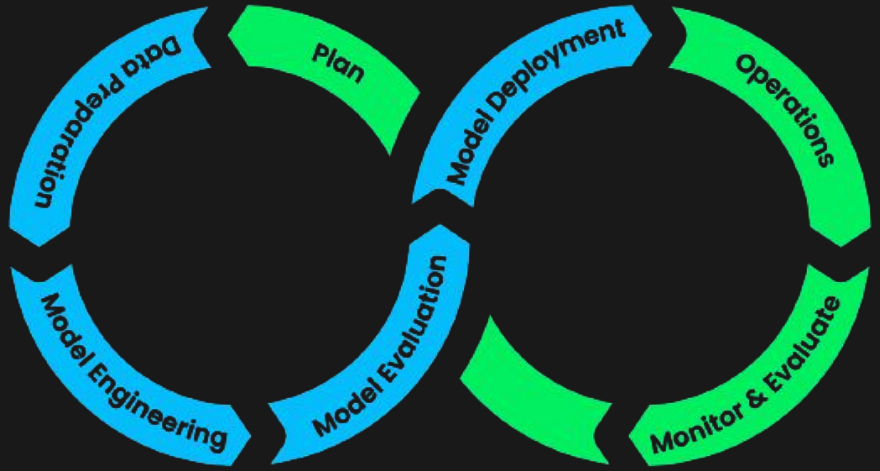
Accurate credit risk assessment can lead to fairer terms and lower interest rates for low-risk borrowers. This project can improve the overall customer experience by tailoring loan offers to individual credit profiles.

06
Economic Impact

Reducing loan defaults can have a positive impact on the overall economy by minimizing the disruptions caused by financial crises or banking system instability.

METHODOLOGY

01 ML Lifecycle



METHODOLOGY

02 DATA LOADING

Data loading was the initial step in the data science and machine learning process, where relevant datasets were imported and loaded into the analysis environment. This step was crucial as it provided the foundation for all subsequent data manipulation, analysis, and modeling. Depending on the dataset's format and source, various methods and libraries were used to retrieve and prepare the data for further tasks.

03 EDA

EDA, which stands for Exploratory Data Analysis, was essential for making informed decisions throughout the data science lifecycle. It involved techniques such as data visualization, summary statistics, and data manipulation to uncover meaningful information from the data.

METHODOLOGY

04 FEATURE SELECTION

During this phase, relevant features were chosen from the dataset, and less important or redundant ones were excluded. This process helped streamline the model and improve its performance, as it reduced dimensionality and potential noise in the data. Techniques like statistical tests, correlation analysis, and domain expertise were applied to identify the most informative features for analysis and modeling.

05 Oversampling using ROSE for class imbalance

ROSE, or Random Over-Sampling Examples, was used to combat class imbalance in datasets. It achieved balance by creating synthetic samples for the minority class through interpolation. This approach bolstered the model's performance in tasks like fraud detection, and credit risk assessment when facing imbalanced data distributions.

Oversampling using ROSE for class imbalance

METHODOLOGY

06 MODEL TRAINING

In this phase, machine learning algorithms were trained using historical data to learn patterns and relationships. Parameters were optimized to enhance predictive performance. Models were employed for classification, regression, or clustering, aiding decision-making. The choice of algorithms and techniques varied based on the specific problem and dataset.

07 Prediction

During prediction, trained ML models were applied to new data, leveraging learned patterns for decision-making in areas like recommendations, risk assessment, and anomaly detection. Accuracy and reliability were pivotal for model success in its designated task.

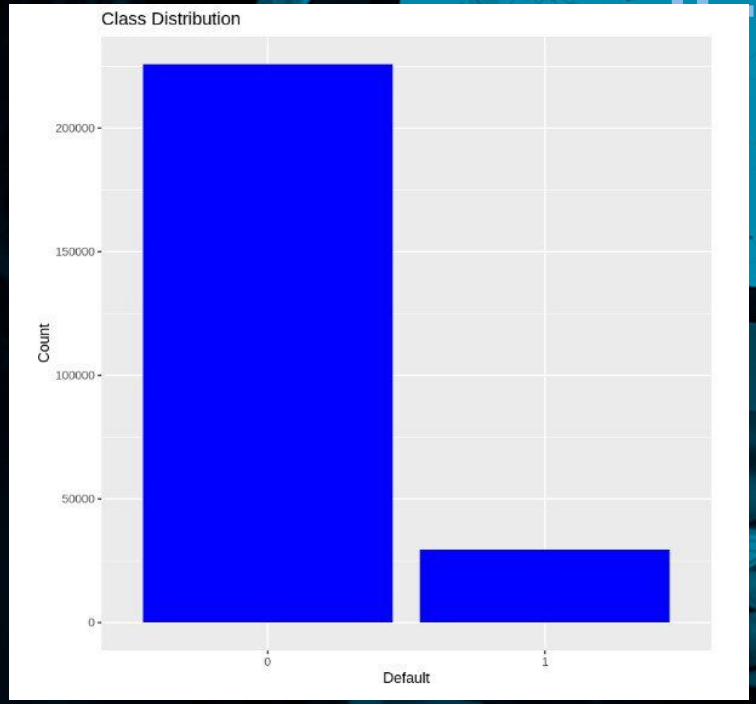
DATASET USED

Features

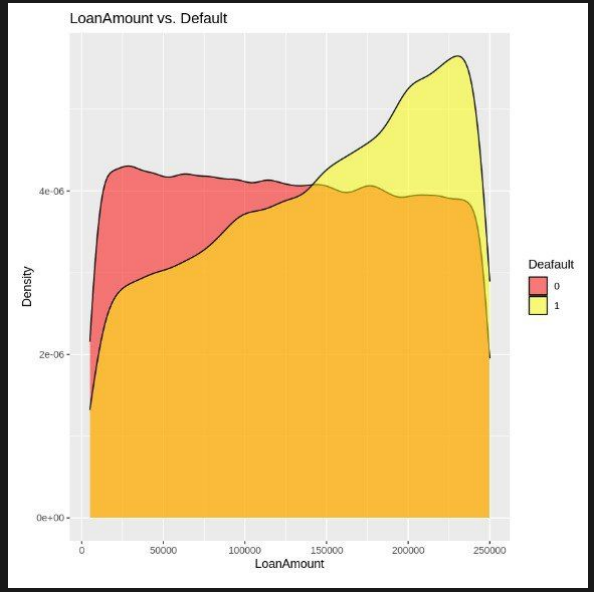
	Column_name	Column_type	Data_type	Description
0	LoanID	Identifier	string	A unique identifier for each loan.
1	Age	Feature	integer	The age of the borrower.
2	Income	Feature	integer	The annual income of the borrower.
3	LoanAmount	Feature	integer	The amount of money being borrowed.
4	CreditScore	Feature	integer	The credit score of the borrower, indicating their creditworthiness.
5	MonthsEmployed	Feature	integer	The number of months the borrower has been employed.
6	NumCreditLines	Feature	integer	The number of credit lines the borrower has open.
7	InterestRate	Feature	float	The interest rate for the loan.
8	LoanTerm	Feature	integer	The term length of the loan in months.
9	DTIRatio	Feature	float	The Debt-to-Income ratio, indicating the borrower's debt compared to their income.
10	Education	Feature	string	The highest level of education attained by the borrower (PhD, Master's, Bachelor's, High School).
11	EmploymentType	Feature	string	The type of employment status of the borrower (Full-time, Part-time, Self-employed, Unemployed).
12	MaritalStatus	Feature	string	The marital status of the borrower (Single, Married, Divorced).
13	HasMortgage	Feature	string	Whether the borrower has a mortgage (Yes or No).
14	HasDependents	Feature	string	Whether the borrower has dependents (Yes or No).
15	LoanPurpose	Feature	string	The purpose of the loan (Home, Auto, Education, Business, Other).
16	HasCoSigner	Feature	string	Whether the loan has a co-signer (Yes or No).
17	Default	Target	integer	The binary target variable indicating whether the loan defaulted (1) or not (0).

This dataset has been taken from Coursera's Loan Default Prediction Challenge and provided us the opportunity to tackle one of the most industry-relevant machine learning problems with a unique dataset that puts our modeling skills to the test. The dataset contains 255,347 rows and 18 columns in total.

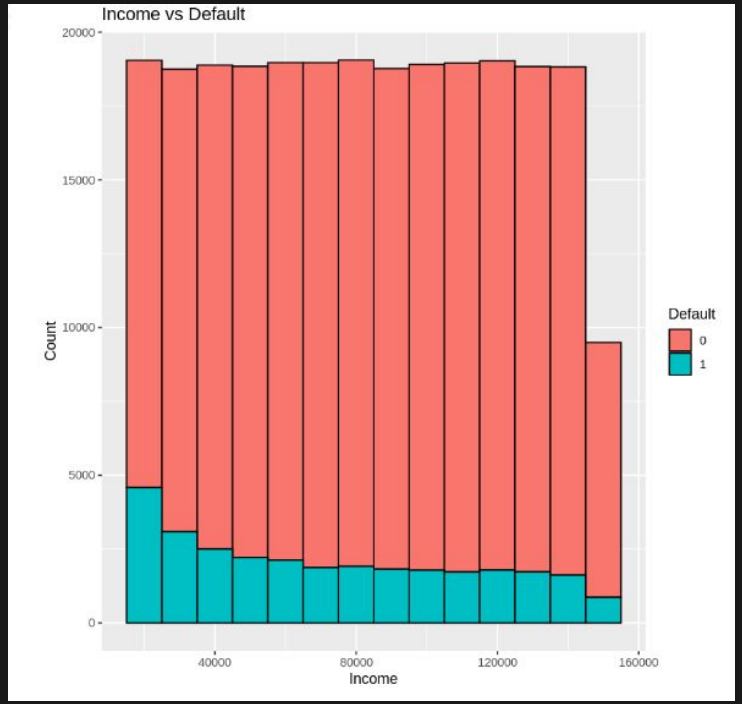
EXPLORATORY DATA ANALYSIS (EDA)



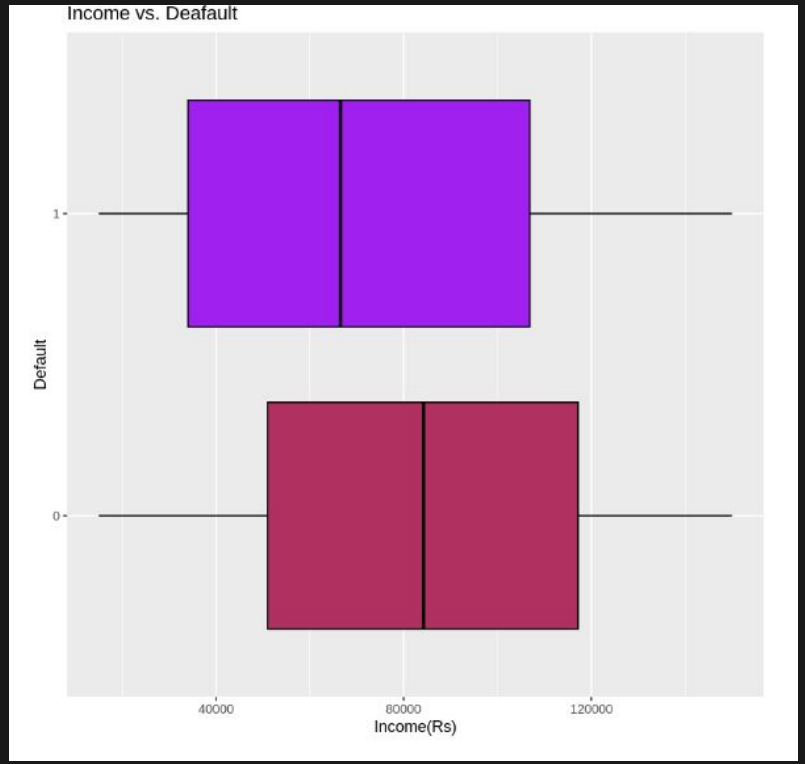
The bar plot displays the class imbalance in our predictor “Default”, which is a concerning issue, and we handled it with the help of oversampling technique, called ROSE (Rank Ordering of Super-Enhancers)



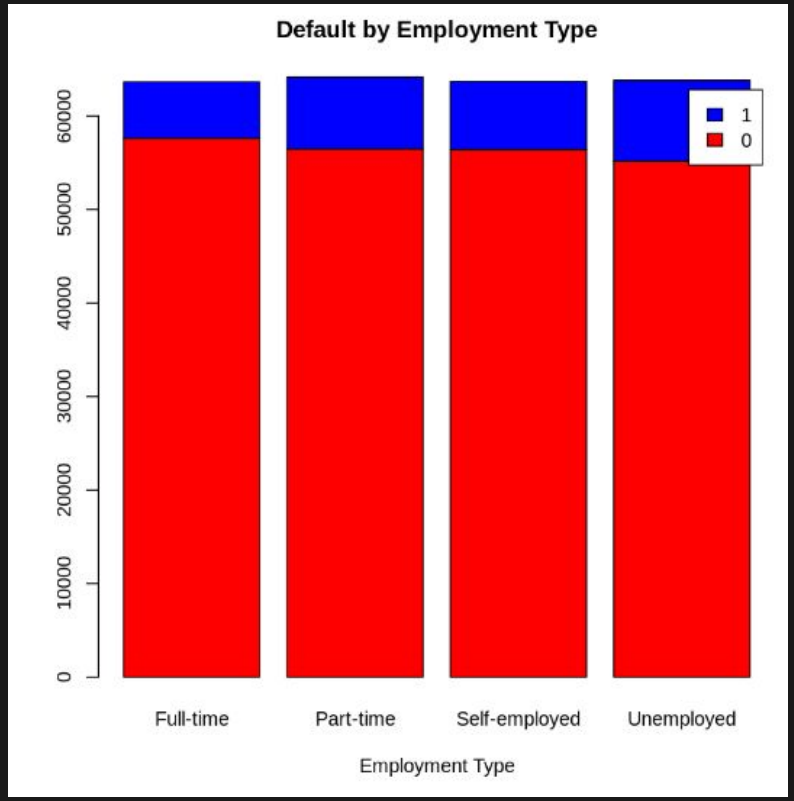
Density Plot



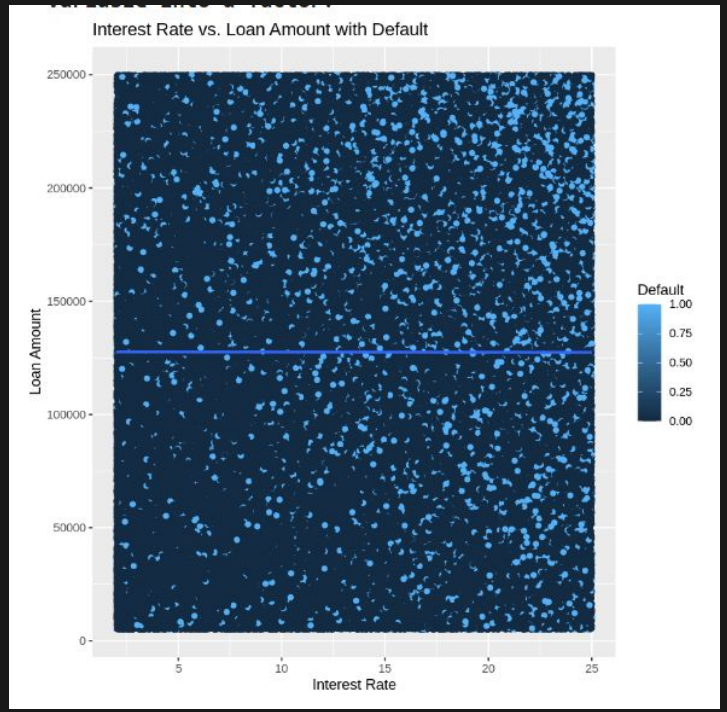
Income vs default



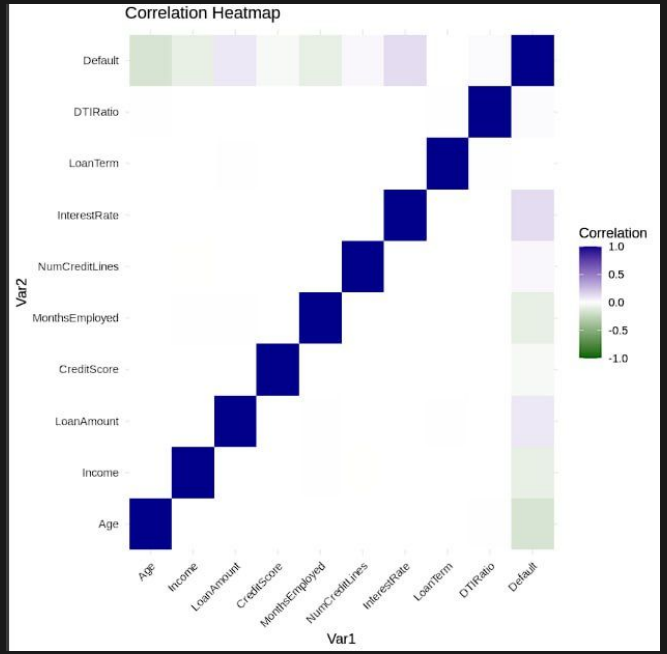
Income vs default


Default by Employment Type
60000
50000
40000
30000
20000
10000
0
1
0
Full-time
Part-time
Self-employed
Unemployed
Employment Type

Bar Plot



Scatter plot



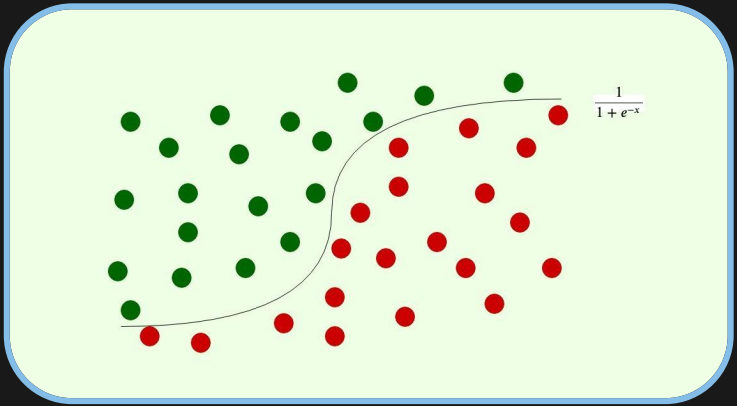
Correlation Heatmap

ALGORITHMS

A diverse set of machine learning algorithms were applied to various data analysis and modeling tasks. These algorithms include:

- Logistic Regression
- Decision Tree
- KNN

Logistic Regression



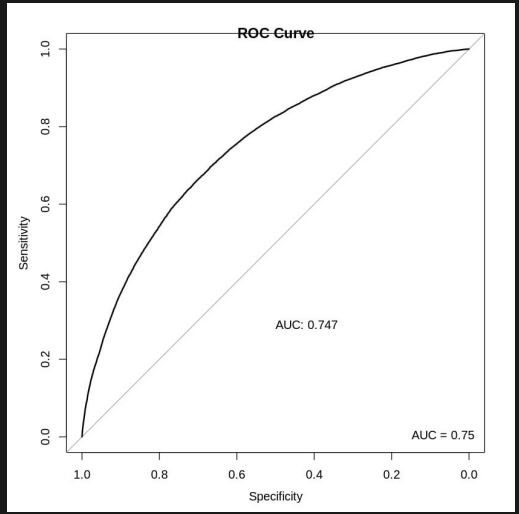
In this project, a logistic regression model was employed as a fundamental machine learning approach for predicting the binary outcome of loan default (1 for default, 0 for non-default). Logistic regression is a widely-used statistical method for binary classification tasks and provides insights into the probability of an event occurring.

Logistic Regression

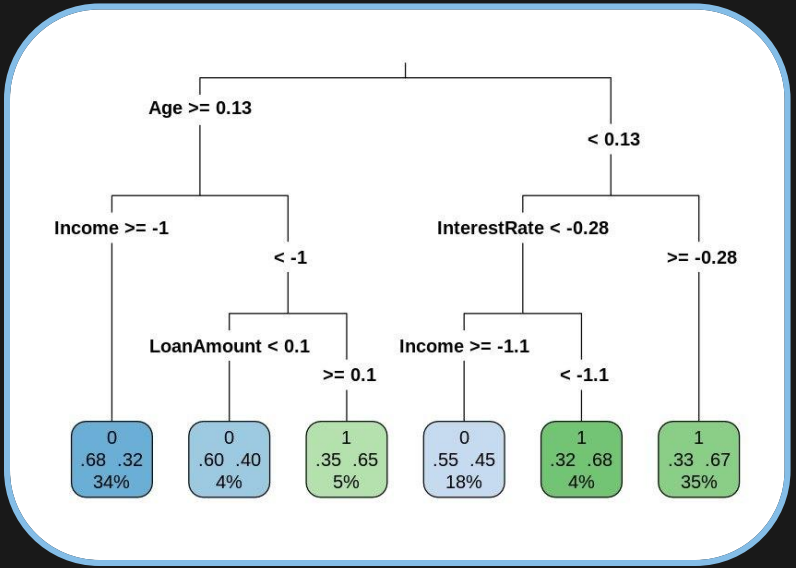
The choice of logistic regression was driven by its simplicity, interpretability, and applicability to binary classification problems. Logistic regression models the relationship between a binary dependent variable (loan default) and one or more independent variables (features) using the logistic function.

Model hyperparameters used: `family: binomial`

Accuracy: `0.49`



Decision Tree



A decision tree is a popular machine learning algorithm used for both classification and regression tasks. It operates by recursively splitting the dataset into subsets based on the most significant feature at each node, eventually creating a tree-like structure of decisions. Decision trees are easy to interpret, making them valuable for understanding the factors influencing a prediction.

In this project, a decision tree classifier was employed as one of the machine learning models to predict the binary outcome of loan default (1 for default, 0 for non-default). The decision tree is a popular algorithm for classification tasks, and it is well-suited for its ability to represent complex decision boundaries and provide interpretability.

Model hyperparameters used: `minsplit = 4, maxdepth = 10, minbucket=4, criterion = "gini", splitter='best', cp=0`

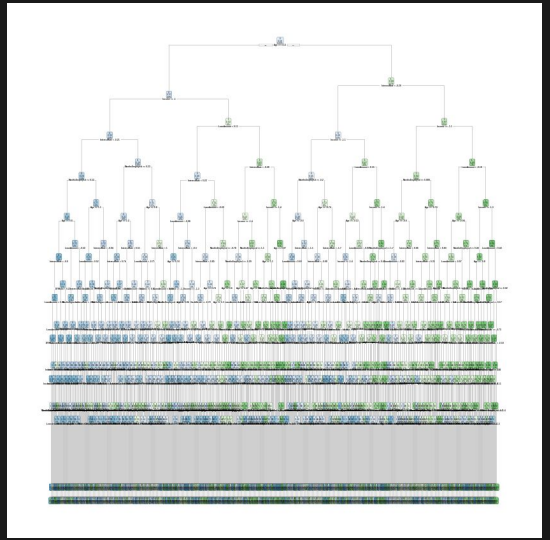
Accuracy: `0.688881`

Confusion matrix:

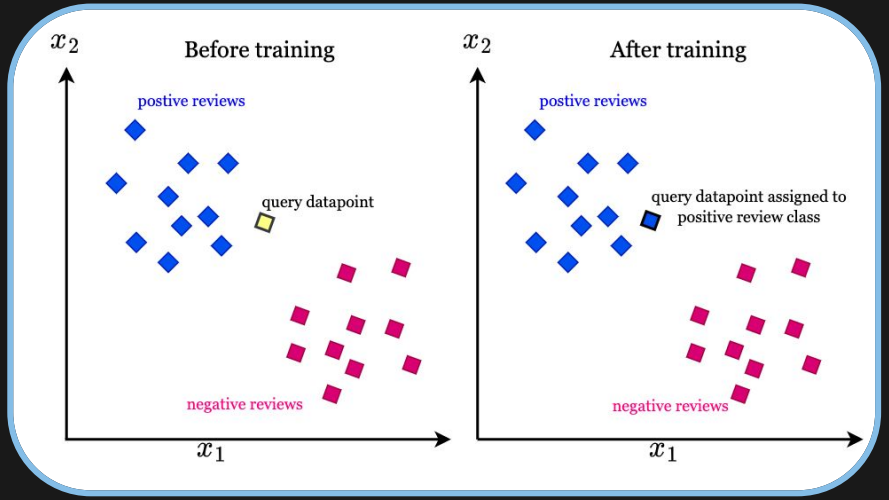
Actual	Predicted 0	1
0	30647	14491
1	13613	31581

Scores:

```
Precision: 0.6854706
Recall: 0.6987874
F1 Score: 0.692065
```



KNN



K-Nearest Neighbors (KNN) is a non-parametric, instance-based machine learning algorithm used for classification tasks. The KNN algorithm classifies data points based on their proximity to other data points in a feature space. This model is suitable for a wide range of classification tasks, including predicting loan defaults.

KNN

The KNN algorithm was chosen for this project due to its simplicity and effectiveness, particularly in cases where decision boundaries are non-linear. KNN is a versatile algorithm that can be adapted to various datasets and is especially valuable when dealing with multivariate data.

Model hyperparameters used: `k <- 3, distance = "Manhattan"`

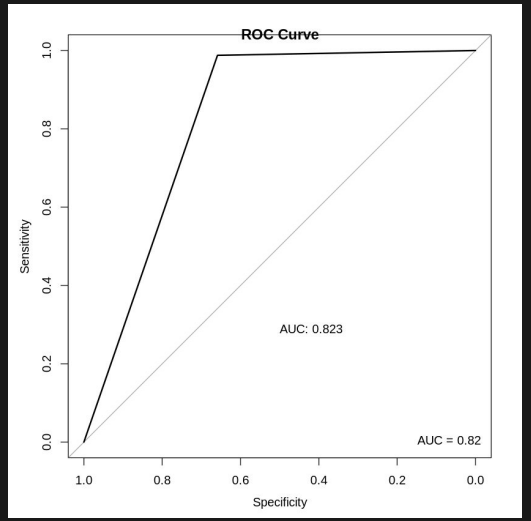
Accuracy: `0.8231709`

Confusion matrix:

```
        Predicted
Actual      0     1
     0  29737 15401
     1    555 44541
```

Scores:

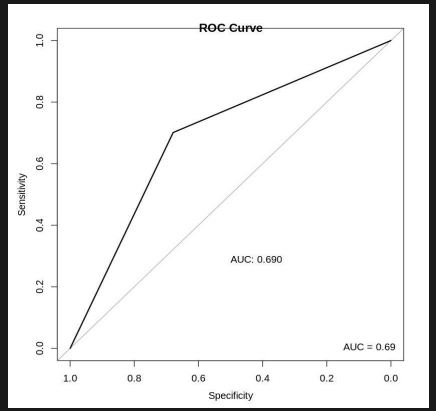
```
Precision: 0.7430683
Recall: 0.9876929
F1 Score: 0.8480931
```



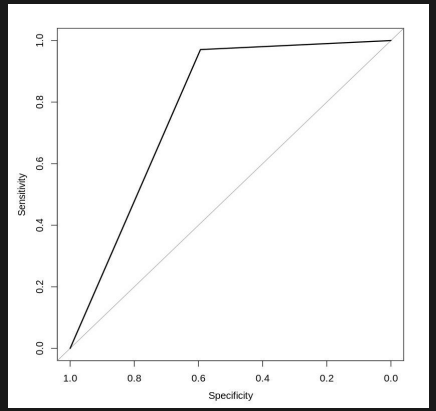
RESULTS

The best performing models were KNN with an accuracy of **82%** followed by Decision Trees with an accuracy of **69%**

Decision Tree



KNN



Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.49	-	-	-
Decision Trees	0.69	0.68	0.71	0.69
KNN	0.82	0.74	0.99	0.85

REFERENCES

- Nikhil. (2021), "Loan Default Prediction Dataset," Kaggle, Available: https://www.kaggle.com/datasets/nikhil1e9/loan-default/data
- Xu, J., Lu, Z., & Xie, Y. "Loan default prediction of Chinese P2P market: a machine learning methodology." Scientific Reports, vol. 11, no. 1, 2021, Article ID 18759, pp. 1-8.
- Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. "Loan default prediction using decision trees and random forest: A comparative study." In IOP Conference Series: Materials Science and Engineering, vol. 1022, no. 1, p. 012042. IOP Publishing, 2021.
- Bhatia, S., Sharma, P., Burman, R., Hazari, S., & Hande, R. "Credit scoring using machine learning techniques." International Journal of Computer Applications, vol. 161, no. 11, 2017, pp. 1-4.
- Addo, P. M., Guegan, D., & Hassani, B. "Credit risk analysis using machine and deep learning models." Risks, vol. 6, no. 2, 2018, pp. 38.
- Bracke, P., Datta, A., Jung, C., & Sen, S. "Machine learning explainability in finance: an application to default risk analysis." , 2019.
- Kadam, A., Pawar, H., Phapale, S., & Ganeshe, C. "Prediction of loan defaulter using machine learning - IJCRT." IJCRT, December 12, 2021. [URL: https://www.ijcrt.org/papers/IJCRT2112549.pdf]

Thank you!