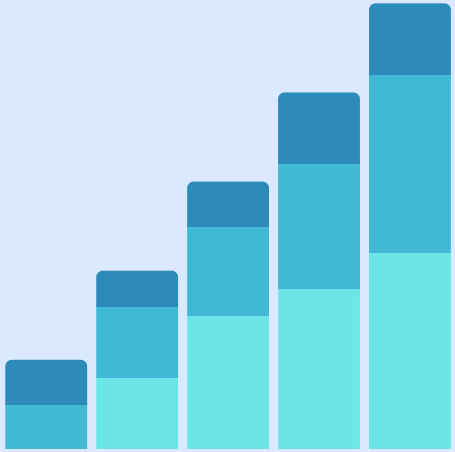
LOAN DEFAULT PREDICTION AND RISK ANALYSIS

MURAD VALIYEV
JAVIDAN HAJIYEV

24 Jan 2025

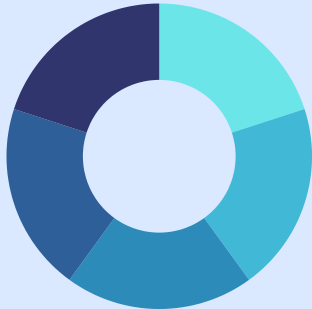
ABSTRACT



This project predicts loan defaults using a dataset with features like age, income, and debt. We built a logistic regression model to classify applicants and found that income, debt-to-income ratio, and employment history significantly impact default risk. The model performed well in predicting defaults.

MOTIVATION

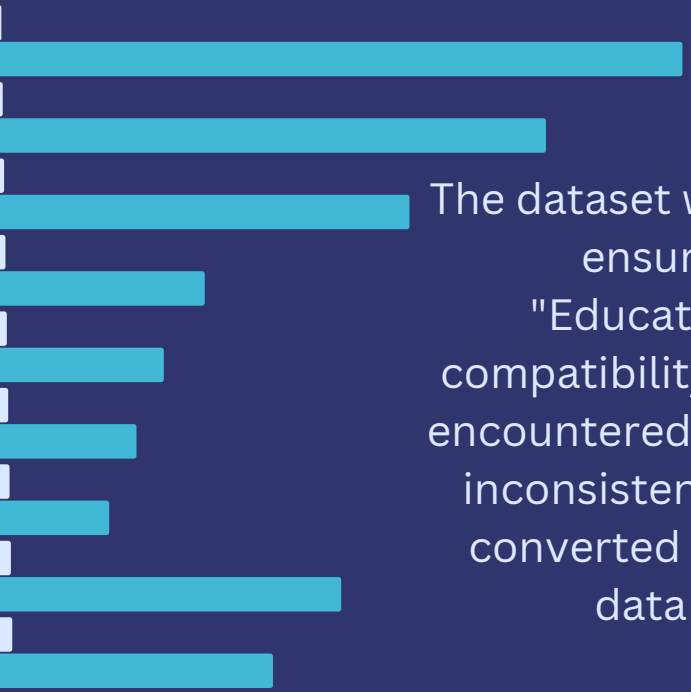
THE GOAL IS TO PREDICT LOAN DEFAULTS, HELPING LENDERS MAKE INFORMED DECISIONS, REDUCE FINANCIAL LOSSES, AND OFFER TAILORED PRODUCTS. ACCURATE PREDICTIONS CAN MITIGATE ECONOMIC CONSEQUENCES FOR BOTH LENDERS AND BORROWERS. UNDERSTANDING DEFAULT RISK FACTORS IMPROVES RISK MANAGEMENT AND HELPS BORROWERS MANAGE BEHAVIORS INFLUENCING LOAN ELIGIBILITY. PREDICTIVE MODELS ENHANCE DECISION-MAKING FOR BETTER FINANCIAL OUTCOMES.



DATASET

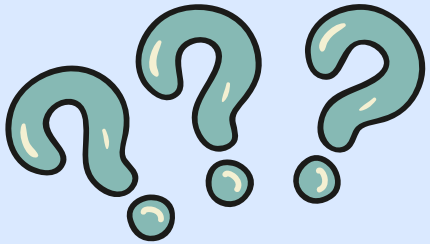
Processed_LoanData.csv Age	Processed_LoanData.csv Ed	Processed_LoanData.csv Employ	Processed_LoanData.csv Address	Processed_LoanData.csv Income	Processed_LoanData.csv Debtinc	Processed_LoanData.csv Creddebt	Processed_LoanData.csv Othdebt	Processed_LoanData Default
41.000	2.00000	5	5	25.000	10.2000	0.3927	2.15730	0.00
24.000	1.00000	3	4	19.000	24.4000	1.3583	3.27765	1.00
36.000	1.00000	0	13	25.000	19.7000	2.7777	2.14730	0.00
27.000	1.00000	0	1	16.000	1.7000	0.1825	0.08949	0.00
25.000	1.00000	4	0	23.000	5.2000	0.2524	0.94364	0.00
52.000	1.00000	24	14	64.000	10.0000	3.9296	2.47040	0.00
37.000	1.00000	6	9	29.000	16.3000	1.7159	3.01110	0.00
48.000	1.00000	22	15	100.000	9.1000	3.7037	5.39630	0.00

DATA PREPARATION AND CLEANING



The dataset was cleaned by removing rows with missing values and ensuring proper formatting. The categorical variable "Education_Level" was encoded into numerical values for compatibility with machine learning models. No major issues were encountered, but data types were checked for consistency and any inconsistencies were handled. The "Default_Status" column was converted to binary values for classification. Once cleaned, the data was prepared for analysis and model building.

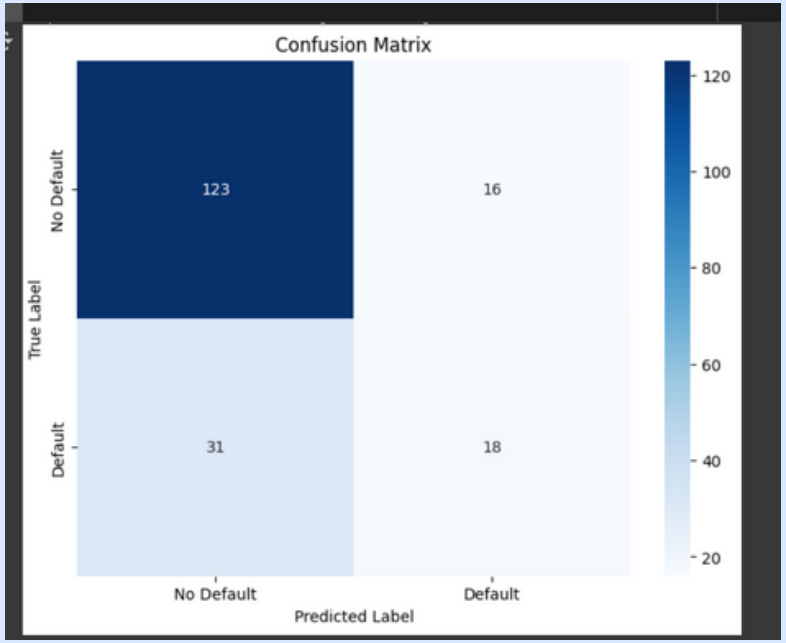
RESEARCH QUESTIONS



- **What factors influence the likelihood of a loan applicant defaulting on a loan?**
- **Can a predictive model be built to accurately classify applicants based on these factors?**

METHODS

- **LOGISTIC REGRESSION:** USED FOR BINARY CLASSIFICATION TO PREDICT THE LIKELIHOOD OF LOAN DEFAULTS BASED ON FINANCIAL AND DEMOGRAPHIC FEATURES.
- **DATA PREPROCESSING:** HANDLED MISSING VALUES, ENCODED CATEGORICAL VARIABLES, AND SCALED NUMERICAL FEATURES TO PREPARE THE DATA FOR ANALYSIS.
- **EVALUATION METRICS:** MODEL PERFORMANCE WAS ASSESSED USING ACCURACY, PRECISION, RECALL, AND ROC CURVE TO EVALUATE CLASSIFICATION EFFECTIVENESS.
- **ASSOCIATION RULES:** APPLIED THE APRIORI ALGORITHM TO IDENTIFY FREQUENT ITEMSETS AND GENERATE ASSOCIATION RULES, EXPLORING RELATIONSHIPS BETWEEN VARIABLES LIKE DEBT LEVELS AND DEFAULT STATUS.

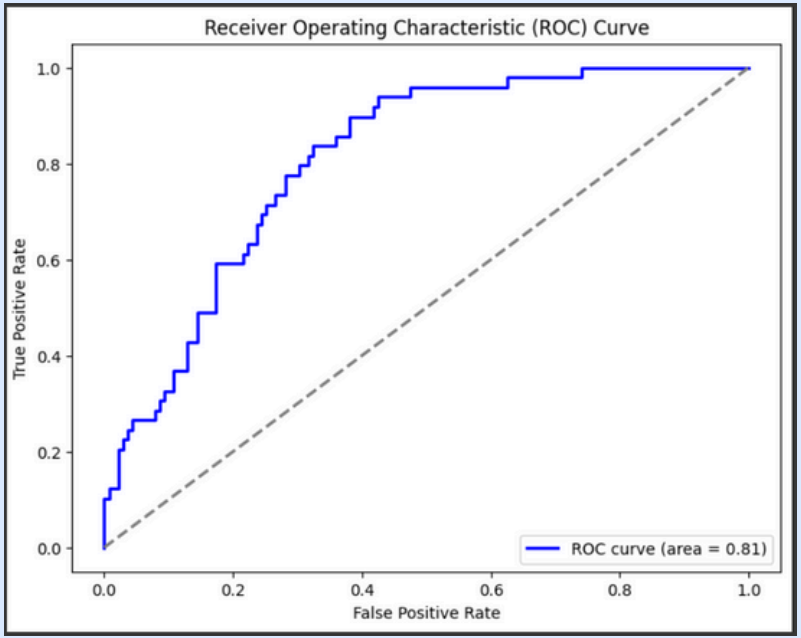


CONFUSION MATRIX

The confusion matrix shows the performance of a classification model by comparing predicted and actual values. It displays true positives, true negatives, false positives, and false negatives, helping to evaluate the model's accuracy and identify errors in predictions.

ROC CURVE

The ROC curve plots the true positive rate against the false positive rate, helping to evaluate a model's performance across different thresholds.
different categories or items.



LIMITATIONS

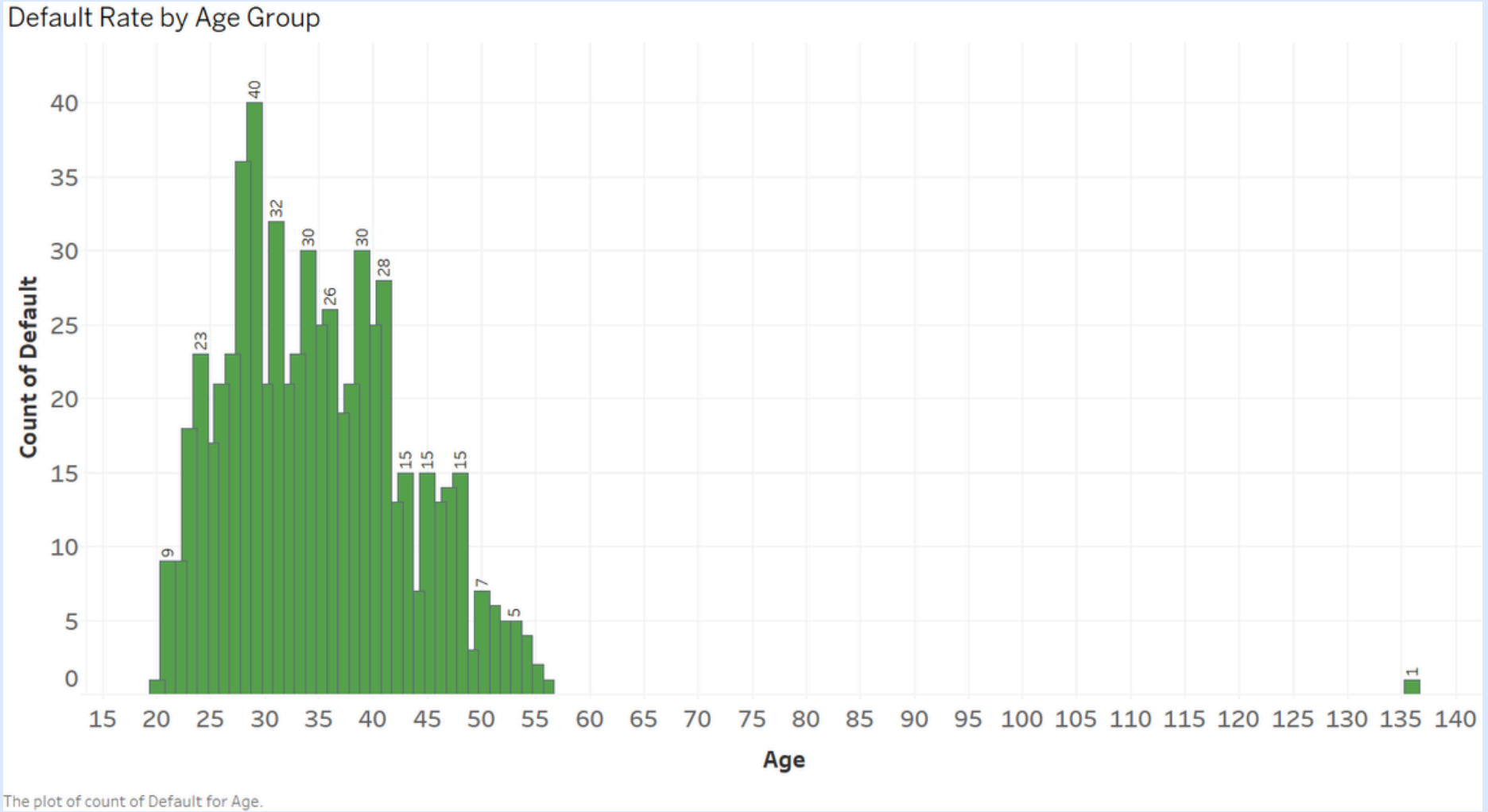
The biggest limitation of this analysis is the relatively small dataset, consisting of only 700 rows. This may impact the generalizability of the model, as a larger dataset could provide more robust insights and improve model accuracy. Additionally, the dataset may not fully capture all factors influencing loan defaults, which could affect the predictive power of the model.

Graphs and Visualizations

1. Income vs. Debt-to-Income Ratio (Scatter Plot)
2. Default Rate by Age Group (Bar Chart)
3. Default Rate Distribution (Pie Chart)
4. Average Income by Education Level (Bar Chart)

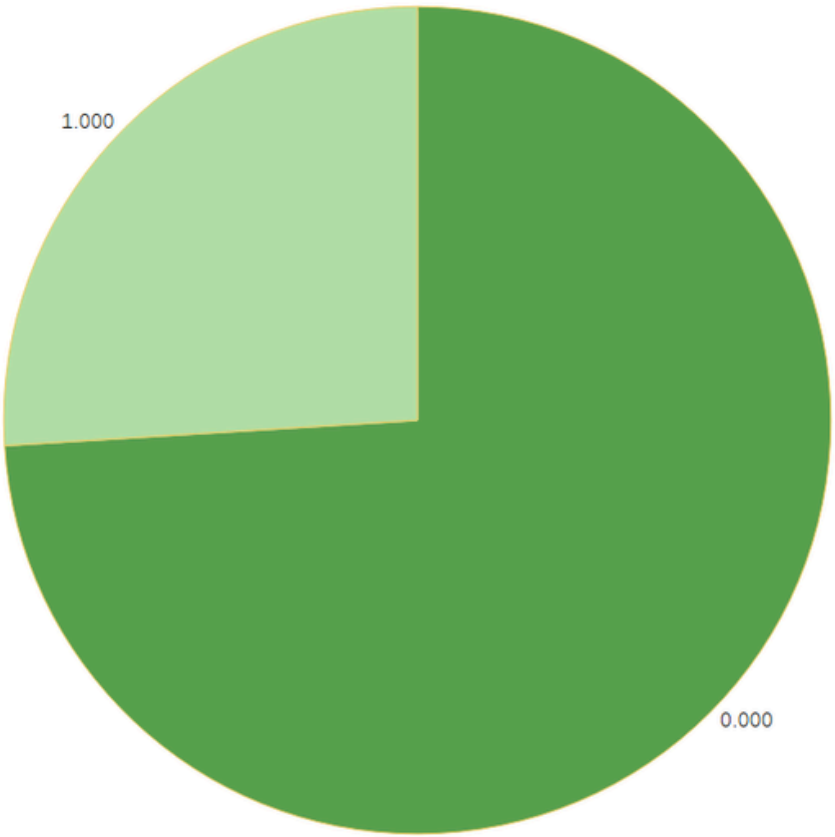
Income vs. Debt-to-Income Ratio

The trend of count of Debtinc for Income. Color shows count of Default.



The plot of count of Default for Age.

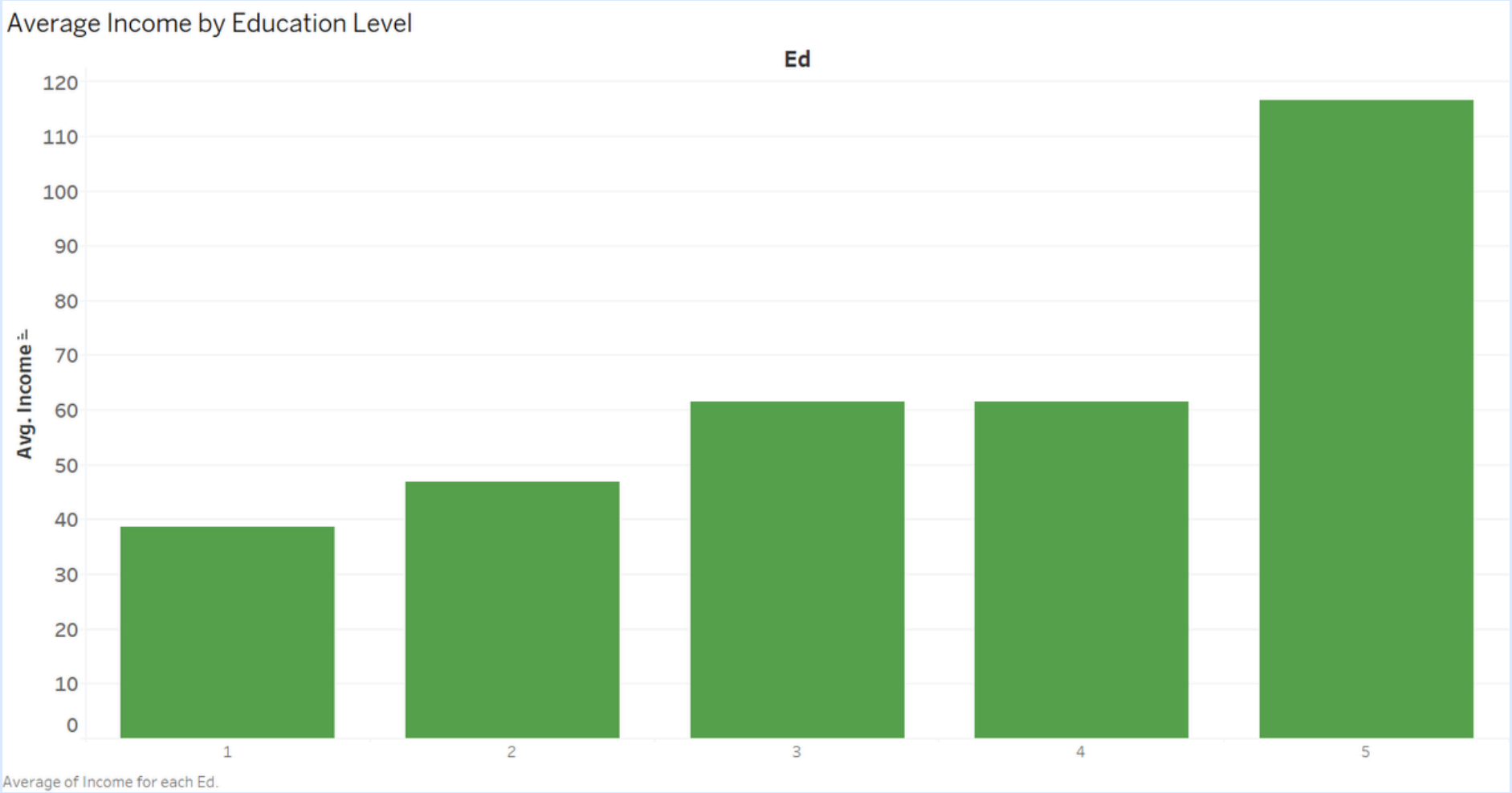
Default Rate Distribution



Count of Default
624

Count of Default
162
462

Default. Color shows details about count of Default. Size shows count of Default. The marks are labeled by Default.


Average Income by Education Level
Ed
Avg. Income
120
110
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
Average of Income for each Ed.

Conclusions

- Key factors influencing loan defaults include income, debt-to-income ratio, and employment history. These variables showed a significant impact on the likelihood of default.
- A logistic regression model was successfully built, performing well in classifying loan defaults. The model's performance was evaluated using classification metrics like accuracy, precision, recall, and ROC curve, indicating a reliable predictive capability.

REFERENCES

DATASET SOURCE

WWW.KAGGLE.COM/DATASETS/MATINMAHMOUDI/LOANS-AND-LIABILITY?RESOURCE=DOWNLOAD&SELECT=LOANDATA_RAW_V1.0.CSV

PANDAS - CLEANING DATA

HTTP://WWW.W3SCHOOLS.COM/PYTHON/PANDAS/PANDAS_CLEANING.ASP