

Predicting Loan Approval Using Classification Models

ISDS 574
Group 1

Aagnya Barot
Vy Trinh
Nam Nguyen
Harshita Manker
Abhinav Dhindsa



AGENDA

I. INTRODUCTION

- Dataset background
- Literature review
- Data Preprocessing

II. PREDICTION METHODS

- Logistic Regression
- kNN
- CART
- Random Forest

III. CONCLUSION

- Key findings & Results
- Comparisons
- Implications
- Limitations
- Future research

DATASET OVERVIEW

 **Source:** Kaggle – [Financial Risk for Loan Approval Dataset](#)

 **Objective:**

Predict whether an applicant is **likely to be approved or denied** for a loan based on their financial and personal attributes.

Target variable: LoanApproved

- 0 = **Loan Denied**
- 1 = **Loan Approved**

DATASET SUMMARY

- **Records:** 20,000

- **Attributes:** 36

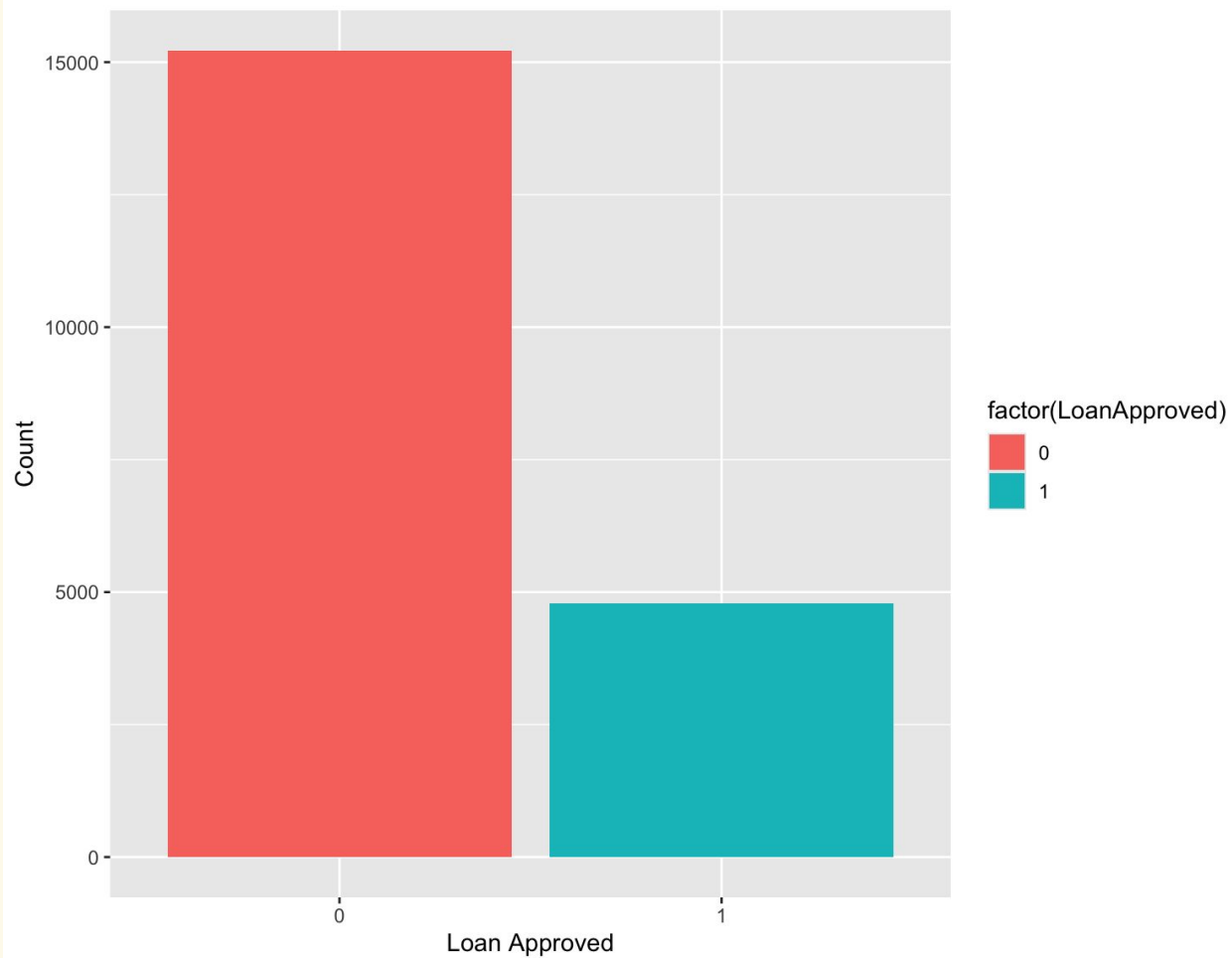
Featuring,

- Demographics
- Credit history
- Employment status
- Income levels
- Existing Debt

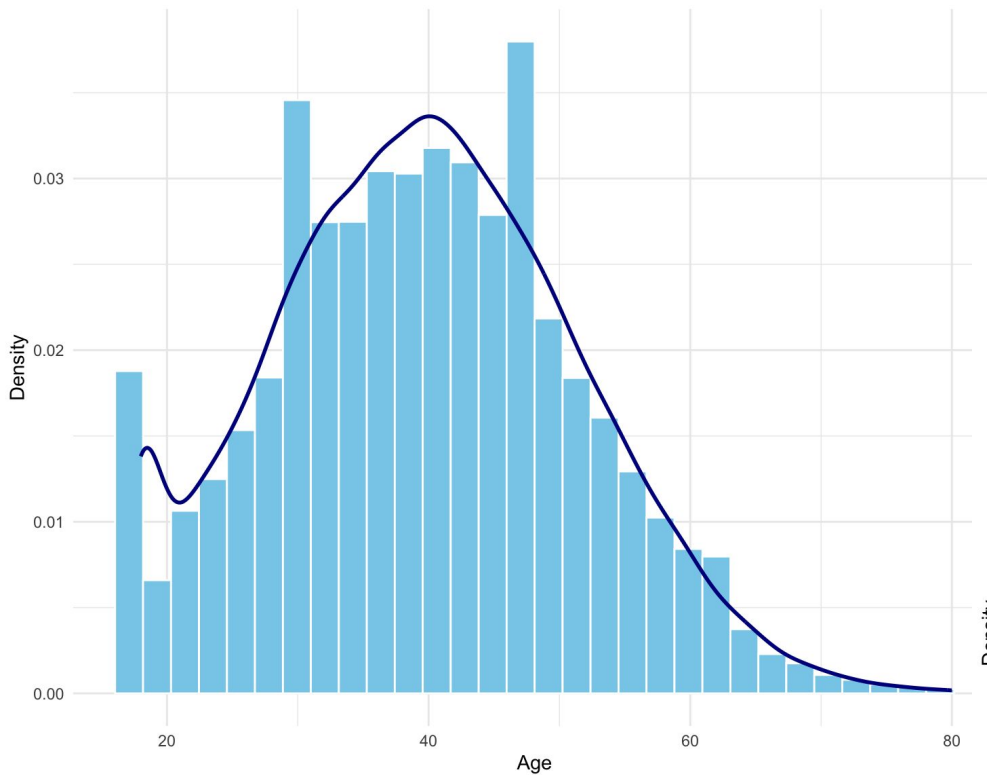
etc,.

Age	AnnualIncome	CreditScore	EmploymentStatus	EducationLevel	Experience	LoanAmount	LoanDuration	MaritalStatus
45	39948	617	Employed	Master	22	13152	48	Married
38	39709	628	Employed	Associate	15	26045	48	Single
47	40724	570	Employed	Bachelor	26	17627	36	Married
58	69084	545	Employed	High School	34	37898	96	Single
37	103264	594	Employed	Associate	17	9184	36	Married
37	178310	626	Self-Employed	Master	16	15433	72	Married
58	51250	564	Employed	High School	39	12741	48	Married
49	97345	516	Employed	High School	23	19634	12	Divorced
34	116841	603	Employed	Bachelor	12	55353	60	Divorced
46	40615	612	Employed	Associate	19	25443	12	Married
34	73646	478	Employed	Associate	10	48716	84	Single
34	15000	591	Employed	Bachelor	11	30088	24	Married
42	74453	573	Employed	Bachelor	21	16154	60	Married
18	100508	580	Employed	Associate	0	20439	60	Married
19	47624	597	Employed	Bachelor	0	27197	36	Married
33	56650	605	Employed	Doctorate	11	12652	36	Single
27	50042	582	Employed	Doctorate	7	19105	60	Single

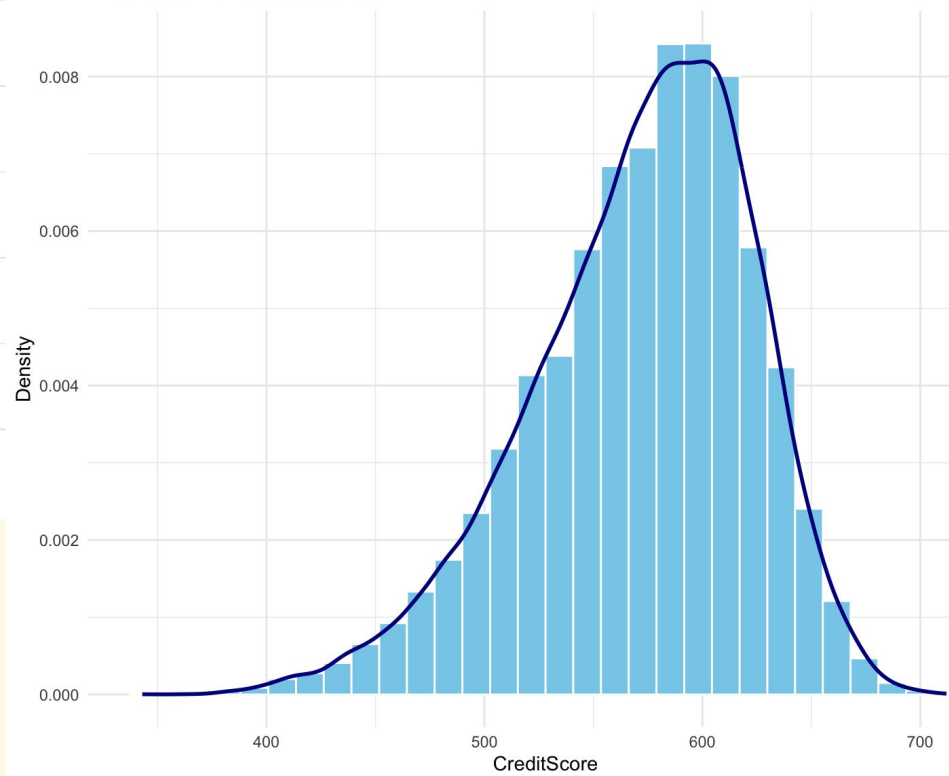
Distribution of Loan Approved

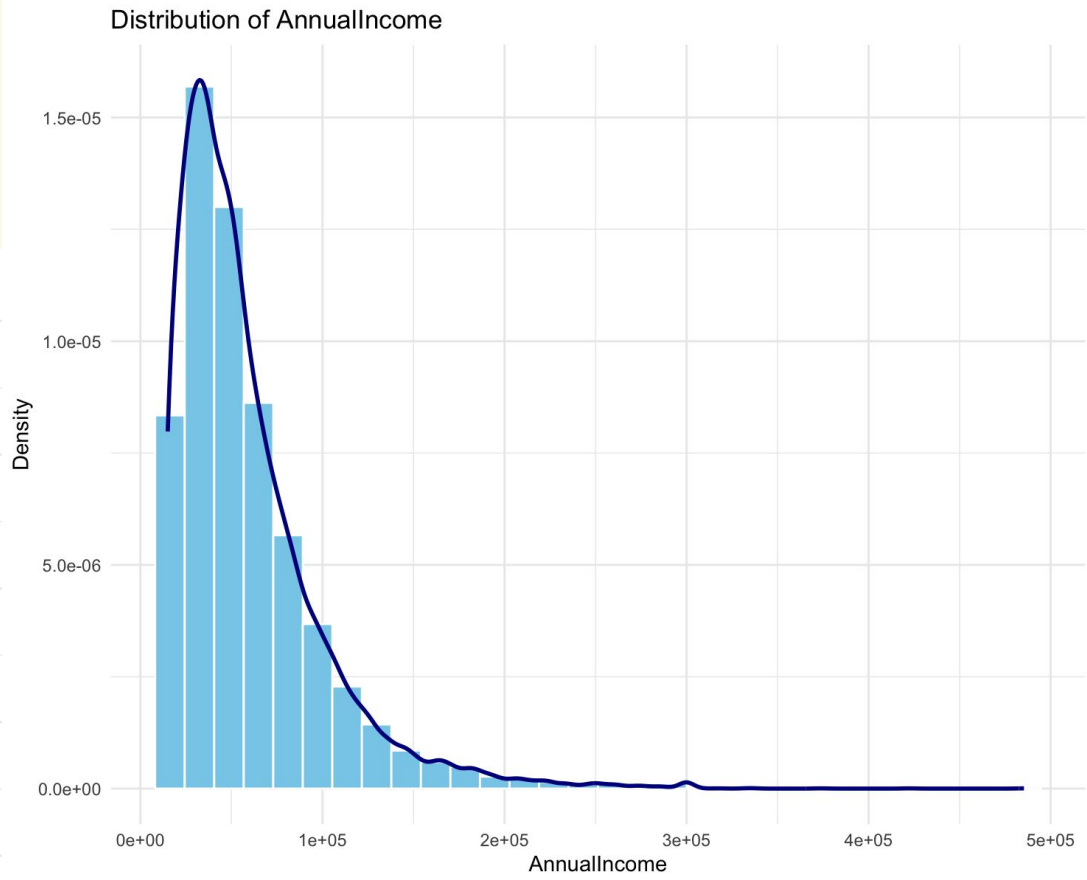
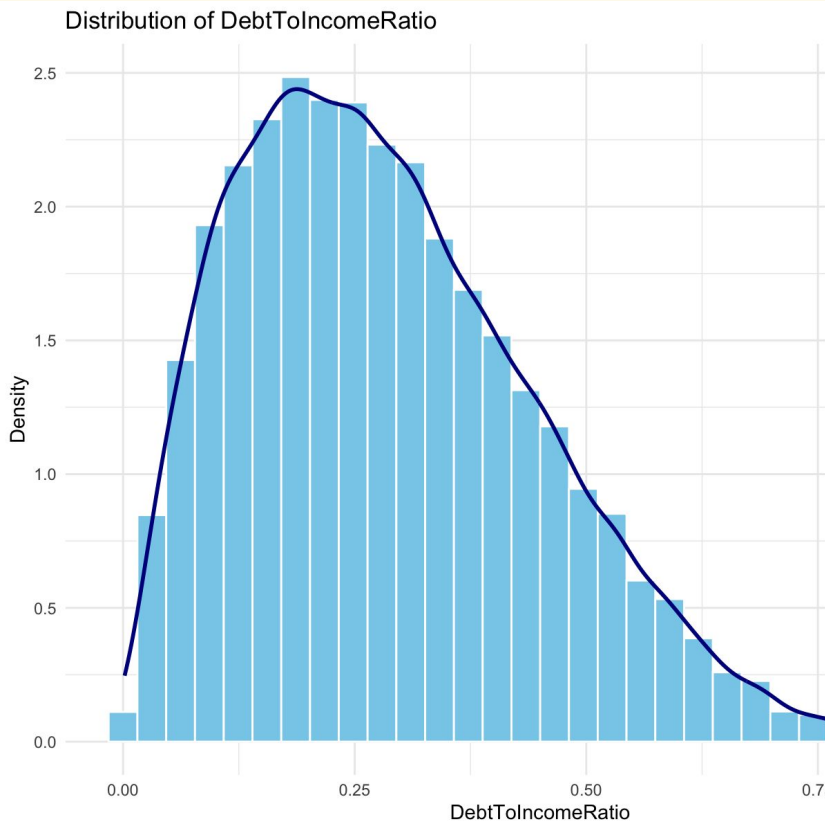


Distribution of Age



Distribution of CreditScore





LITERATURE REVIEW

1. *An Approach for Prediction of Loan Approval using Machine Learning Algorithm*

Authors: M. A. Sheikh, A. K. Goel and T. Kumar

Published: 2020, IEEE Xplore

Overview

Model: Logistic Regression

Focus: Prediction of Loan Approval

Features: Income, age, credit history

Metrics: Accuracy, precision, recall

Relevance to Our Project

Dataset: Kaggle loan approval data

Approach: Binary classification using Logistic Regression

Features: Validates use of structured data (e.g., income, experience)

Insight: The research supports our model choice and evaluation strategy

LITERATURE REVIEW

2. Bank Loan Prediction Using Machine Learning Techniques

Authors: F. M. Ahosanul Haque & Md. Mahedi Hassan

Published: December 2024, American Journal of Industrial and Business Management

Overview

Model: Ada Boosting, Gaussian NB, Random Forest Classifier, Decision Tree Classifier and SVM

Focus: Improve loan approval prediction accuracy and efficiency

Metrics: Models were assessed based on their accuracy in predicting loan approval outcomes.

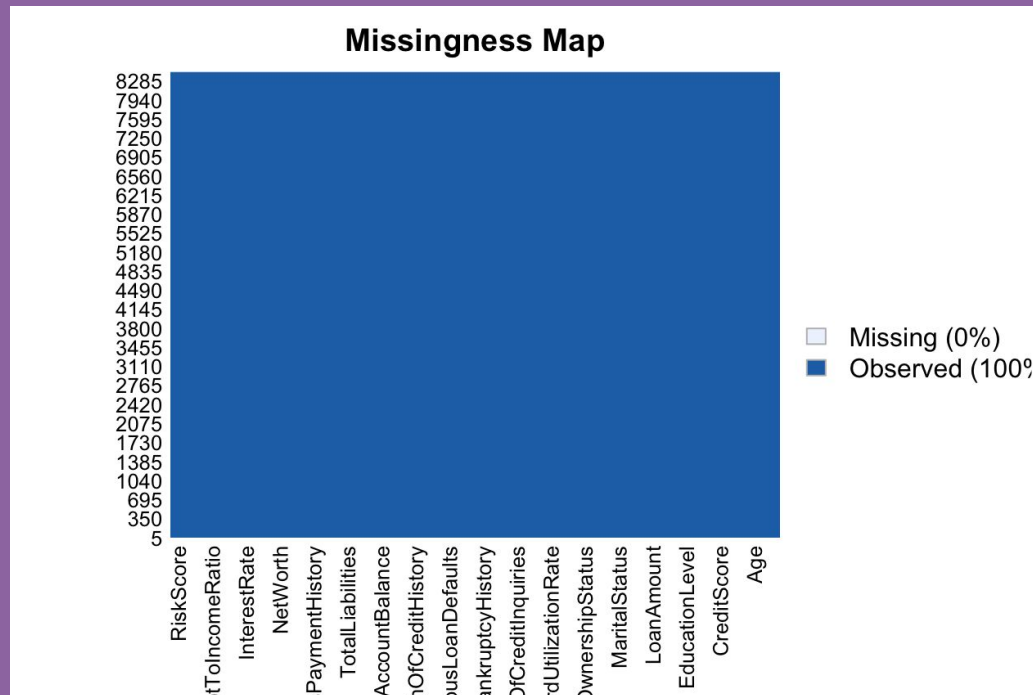
Relevance to Our Project

Random Forest added based on strong results in this study (99.98% accuracy)

Supports use of ensemble methods in our analysis

Data Preprocessing

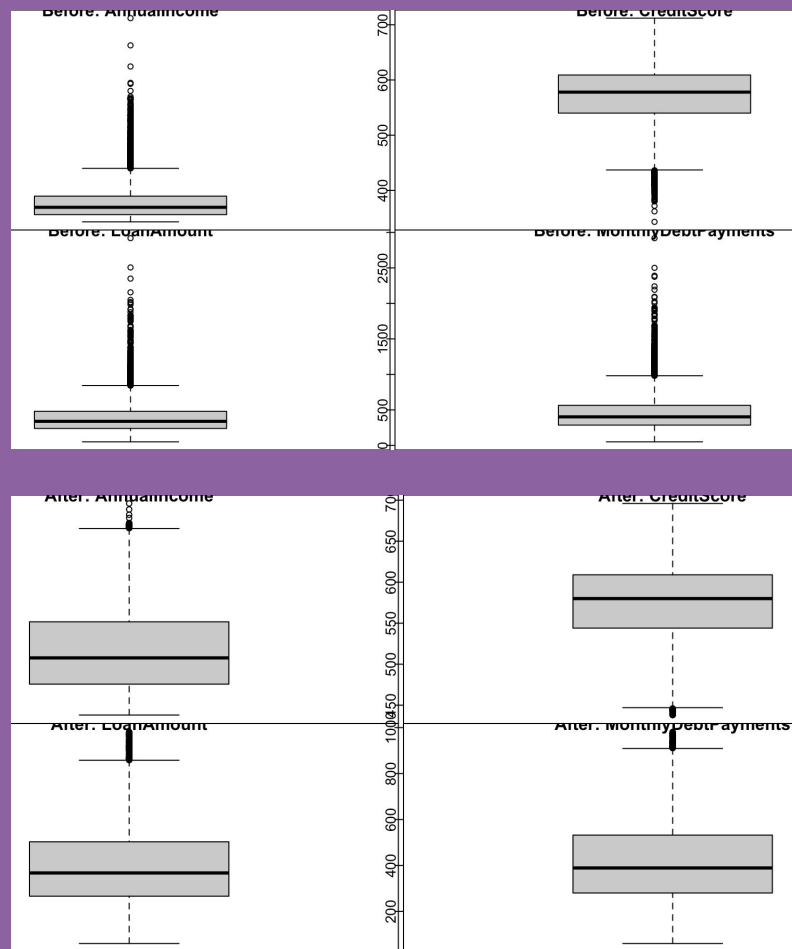
Missing Data Check (Heatmap)



Data Preprocessing

Outliers in Continuous Variables

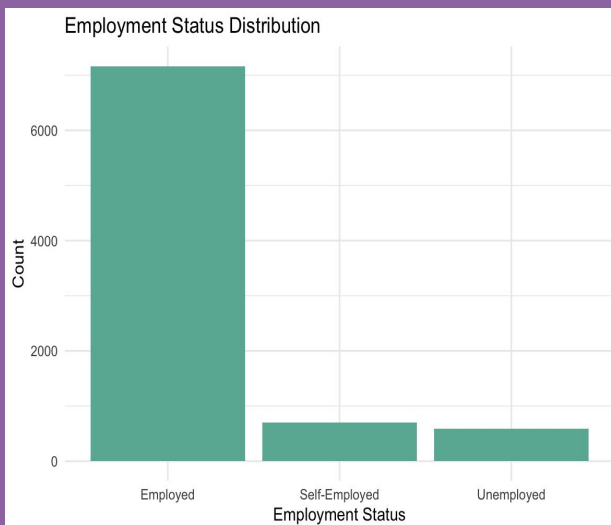
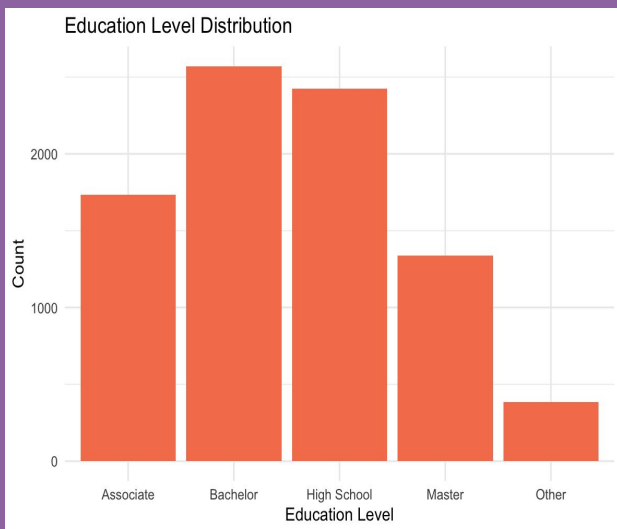
- Removed outliers from all continuous variables using IQR method
- ▼ Rows reduced from 20,000 to 8,450



Data Preprocessing

Categorical Cleanup (Frequency Tables):

Rare levels (<5%) were grouped as “Other” to reduce dimensionality.



Data Preprocessing

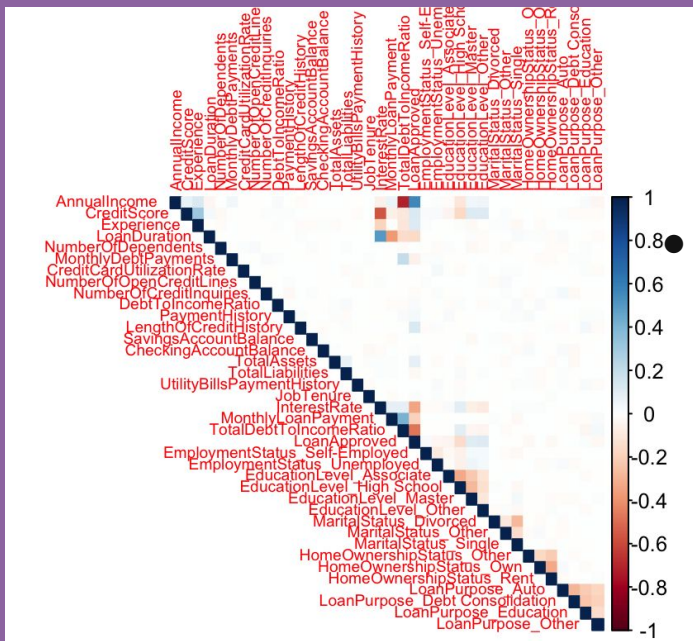
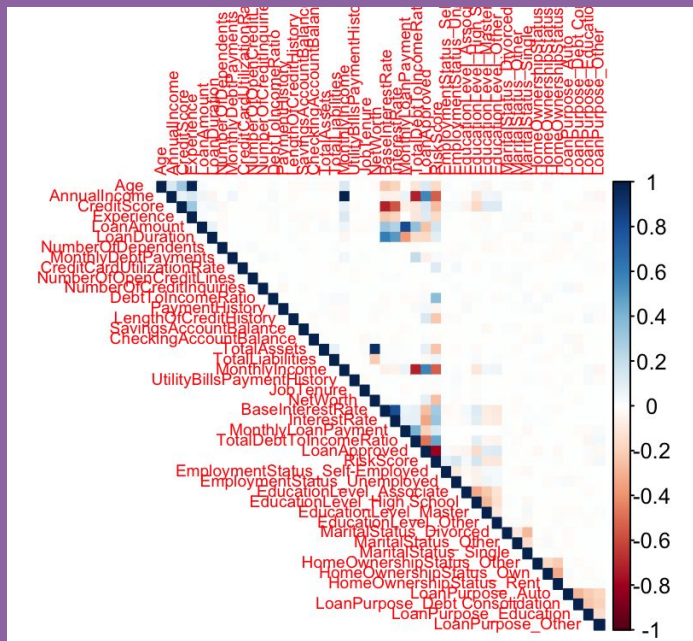
Dummy Variable Encoding- Created dummy variables for 5 categorical features using one-hot encoding.

Original Variable	Dummy Variables Created
EmploymentStatus	EmploymentStatus_Self-Employed, EmploymentStatus_Unemployed (Employed was baseline)
EducationLevel	EducationLevel_Associate, High School, Master, Other (Bachelor was dropped)
MaritalStatus	MaritalStatus_Divorced, Single, Other (Married was baseline)
HomeOwnershipStatus	HomeOwnershipStatus_Own, Rent, Other (Mortgage was dropped)
LoanPurpose	LoanPurpose_Auto, Education, Home, Other (Debt Consolidation was baseline)

Data Preprocessing:

Correlation & Multicollinearity Removal

- Removed 6 variables with correlation > 0.8 to address multicollinearity.



Variables removed:
RiskScore,
BaseInterestRate,
MonthlyIncome,
Age, LoanAmount,
NetWorth

Logistic Regression: Predicting Loan Approval

- **Goal:** Predict the categorical outcome (LoanApproved: 0 = No, 1 = Yes)
- **Variable Selection Methods:** Forward selection, Backward elimination, Stepwise selection
- **Variables Used:** AnnualIncome, CreditScore, MonthlyDebtPayments, CreditCardUtilizationRate, DebtToIncomeRatio, TotalDebtToIncomeRatio, BankruptcyHistory, PreviousLoanDefaults, InterestRate, MonthlyLoanPayment, RiskScore, EmploymentStatus_Employed, LoanDuration
- **Selection Method Used:** Backward Selection
- **Cutoff points analyzed:** 0.5, 0.3, 0.1

Logistic Regression: Model Evaluation at Multiple Cutoffs

Cutoff	Accuracy	Sensitivity	Specificity
0.5	0.953	0.908	0.966
0.3	0.951	0.955	0.95
0.1	0.927	0.984	0.91

Logistic Regression: Results & Insights

Predictor	OR	p-value	Interpretation
InterestRate	~0	< 0.001	↑ InterestRate → ↓ Approval Odds
AnnualIncome	1	< 0.001	↑ Income → Slight ↑ Approval Odds
CreditScore	0.947	< 0.001	↑ CreditScore → ↓ Approval (unexpected)
LengthOfCreditHistory	1.23	< 0.001	Longer history → ↑ Approval Odds
EducationLevel_Master	5.04	< 0.001	Master's degree → ↑ Approval Odds
EmploymentStatus_Unemp	0.03	< 0.001	Unemployed → ↓↓ Approval Odds

- Results based on **Backward Selection Method**.
- **Cutoff** = 0.5 chosen for best balance of sensitivity and specificity.
- **Key predictors**: RiskScore ↓, AnnualIncome ↑, CreditScore ↓, InterestRate ↓, EmploymentStatus_Unemp ↓
- **CreditScore** and **InterestRate** strongly reduce loan approval odds.
- Higher **AnnualIncome** slightly increases loan approval odds.

K-Nearest Neighbour

- **Goal** - To predict whether a loan will be approved using **KNN Classification** algorithm
- Selecting **best-K** value: We have used **cross-validation** to select the best value for K ranging from 1 to 20
- Provides a more reliable estimate of model performance
 - ◆ Helps to avoid overfitting or underfitting
 - ◆ Only odd values for K are used
- KNN has been performed on
 - ◆ All Variables
 - ◆ Variables from Logistic Regression
 - ◆ Screened Variables(variables with a correlation of >0.2 or <-0.2 with the outcome)

KNN: On All Variables

Results	Accuracy	Specificity	Sensitivity
Cutoff 0.5	84.18%	97.24%	39.75%
Cutoff 0.3	84.02%	98.92%	33.33%
Cutoff 0.1	78.81%	88.15%	47.04%

The optimal value of k at cutoff 0.5 is 9

The optimal value of k at cutoff 0.3 is 17

The optimal value of k at cutoff 0.1 is 1

Insights on KNN using All variables

- Higher cutoff reflects better at predicting 0's (non-approvals) but also many missed 1's (Approvals)
- A lower cutoff threshold catches more approved loans (1s), but it also increases the number of false approvals (false positives).
- Unnecessary or noisy variables can hurt model accuracy, especially if they don't help in distinguishing between approved and non-approved loans.
- Although the model provides a decent performance, but lower sensitivity across cutoffs suggest that careful consideration of variable selection improves KNN's reliability.

KNN: On Variables Obtained from Logistic Regression

Model	Accuracy	Sensitivity	Specificity
Stepwise	Cutoff 0.5 = 87.14% Cutoff 0.3 = 87.17% Cutoff 0.1 = 87.14%	Cutoff 0.5 = 51.38% Cutoff 0.3 = 51.56% Cutoff 0.1 = 51.38%	Cutoff 0.5 = 97.65% Cutoff 0.3 = 97.65% Cutoff 0.1 = 97.65%
Forward	Cutoff 0.5 = 86.78% Cutoff 0.3 = 86.78% Cutoff 0.1 = 86.78%	Cutoff 0.5 = 52.25% Cutoff 0.3 = 52.22% Cutoff 0.1 = 52.25%	Cutoff 0.5 = 96.93% Cutoff 0.3 = 96.93% Cutoff 0.1 = 96.93%
Backward	Cutoff 0.5 = 87.14% Cutoff 0.3 = 87.17% Cutoff 0.1 = 87.14%	Cutoff 0.5 = 51.38% Cutoff 0.3 = 51.56% Cutoff 0.1 = 51.38%	Cutoff 0.5 = 97.65% Cutoff 0.3 = 97.65% Cutoff 0.1 = 97.65%

For Stepwise model

The optimal value of k at cutoff 0.5 is 7
The optimal value of k at cutoff 0.3 is 17
The optimal value of k at cutoff 0.1 is 1

For Forward model

The optimal value of k at cutoff 0.5 is 17
The optimal value of k at cutoff 0.3 is 17
The optimal value of k at cutoff 0.1 is 1

For Backward model

The optimal value of k at cutoff 0.5 is 7
The optimal value of k at cutoff 0.3 is 17
The optimal value of k at cutoff 0.1 is 1

Insights from KNN with Logistic Regression variables

- Changing threshold has minimal impact on the performance
- High Specificity indicates strong performance in predicting loan denials, where as low Sensitivity indicates that the model struggles to identify loan approvals
- Dataset is likely to have more loan denials than approvals, biasing KNN towards majority class
- KNN struggles when the classes are not well separated in feature space
- Even after performing logistic regression and using those variables, it can still affect the results as too many predictors can make distance calculations less meaningful

KNN: On Screened Variables

Cutoffs	Accuracy	Sensitivity	Specificity
Cutoff 0.5	89.86%	73.43%	94.69%
Cutoff 0.3	89.86%	73.43%	94.69%
Cutoff 0.1	89.86%	73.43%	94.69%

The optimal value of k at cutoff 0.5 is 19

The optimal value of k at cutoff 0.3 is 17

The optimal value of k at cutoff 0.1 is 1

Insights on KNN using Screened variables

- The KNN model using screened variables delivers the best overall performance.
- The Accuracy is consistent high at 89.86% across all the cutoffs, showing model's stability.
- Sensitivity is higher as compared to other KNN setups indicating strong ability to predict loan approvals.
- Specificity is also strong, meaning the model maintains low false positives.
- Cutoff have no significant impact on the model performance.
- Screened variables create a cleaner and more separable feature space, making KNN less sensitive to threshold changes.
- **Using a carefully selected subset of variables leads to more balanced and reliable KNN model.**

CART

Parameter (default setting)

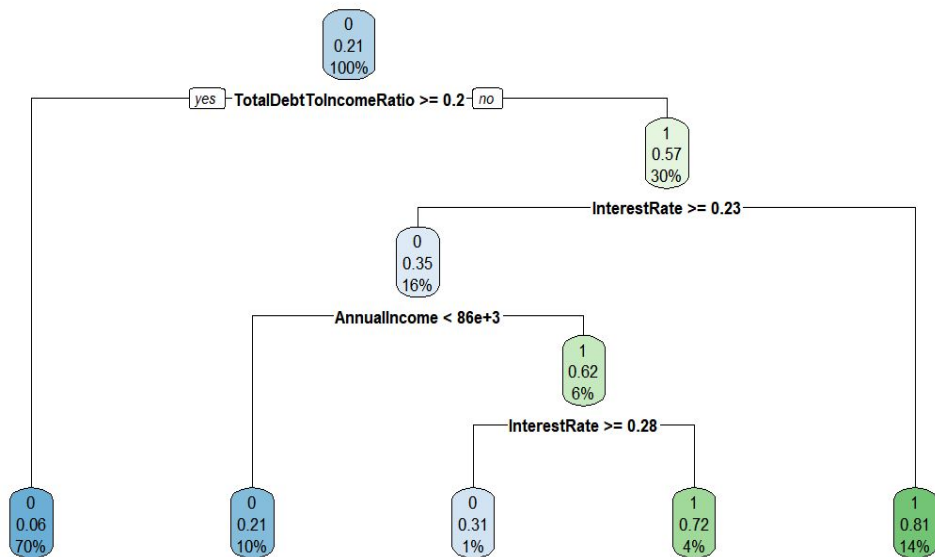
- **Use all of variable in reprocessing data to predict Loan Approval (except Application Date)**
- **Minsplit = 20** : Minimum number of observations exist in a node to split
- **Minbucket = round(minsplit/3)= 7** : number of observations in any terminal
- **Maxdepth = 30** : maximum depth of any node of the final tree

Minimum Error Tree: tree with smallest classification error rate

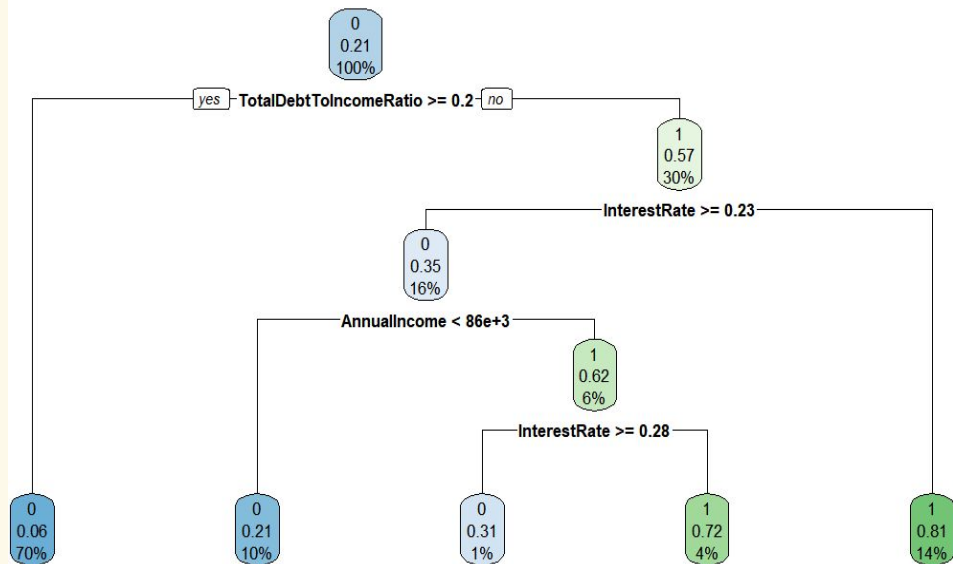
Best Pruned Tree: tree within one standard error of the minimum error tree

CART

Min Error Tree



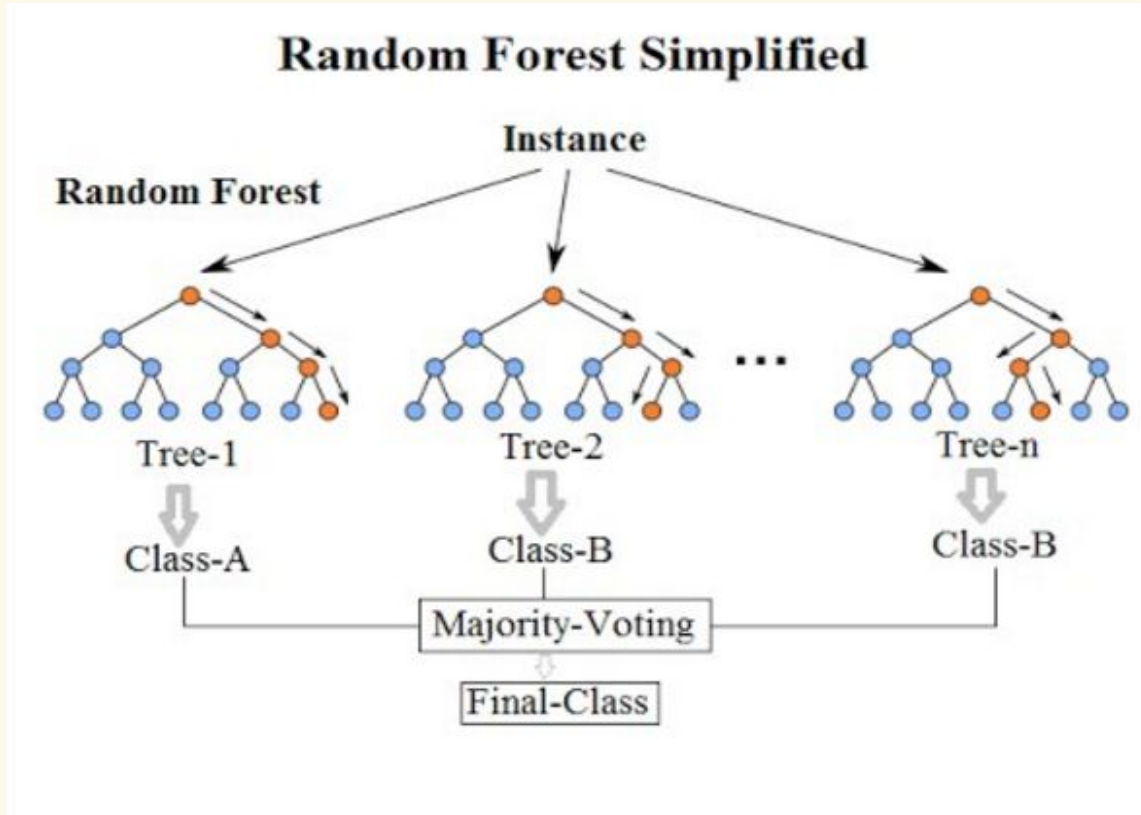
Best Pruned Tree



CART

		Minimum Error Tree	Best Pruned Tree
Cutoff= 0.5	Accuracy Sensitivity Specificity	88.0% 67.5% 93.8%	88.0% 67.5% 93.8%
Cutoff= 0.3	Accuracy Sensitivity Specificity	86.9% 68.0% 92.2%	86.9% 68.0% 92.2%
Cutoff= 0.1	Accuracy Sensitivity Specificity	81.7% 79.8% 82.2%	81.7% 79.8% 82.2%

Random Forest



Random Forest

		Default Setting (mtry= 6, number of tree = 500)
Cutoff= 0.5	Accuracy Sensitivity Specificity	91.6% 74.4% 96.4%
Cutoff= 0.3	Accuracy Sensitivity Specificity	89.7% 89.6% 89.7%
Cutoff= 0.1	Accuracy Sensitivity Specificity	76.6% 98.9% 70.5%

Key Findings and Insights

- **Data preprocessing** is a critical part for a model as it directly impacts the performance. It includes **handling missing values, encoding categorical variables, outlier detection and removal.**
- For our LOAN APPROVAL CLASSIFICATION model, some predictor variables have more influence on the output than others. For ex - **Credit Score, Annual Income, Employment Status, and Interest rate.**
- Scaling the data is not required for **Logistic Regression** but it is necessary for **K-Nearest Neighbour** algorithm as it is **distance based** and scaling helps in improving the accuracy and prediction quality of the model.
- Class Imbalance Affects Model Fairness : If preprocessed data leaves us with more loan rejections than approvals, classification will not be accurate.
- Domain knowledge is equally important to chose relevant features and remove unimportant, misleading variables.

Results Summary

		Logistic Regression	KNN (screened variable)	CART (best pruned tree)	Random Forest (default setting)
Cutoff= 0.5	Accuracy Sensitivity Specificity	95.3% 90.8% 96.6%	89.9% 73.4% 94.7%	88.0% 67.5% 93.8%	91.6% 74.4% 96.4%
Cutoff= 0.3	Accuracy Sensitivity Specificity	95.1% 95.5% 95.0%	89.9% 73.4% 94.7%	86.9% 68.0% 92.2%	89.7% 89.6% 89.7%
Cutoff= 0.1	Accuracy Sensitivity Specificity	92.7% 98.4% 91.0%	89.9% 73.4% 94.7%	81.7% 79.8% 82.2%	76.6% 98.9% 70.5%

Existing Research - Comparison

Nureni and Adekola (2022) found Logistic Regression to be the most accurate in predicting loan defaults, emphasizing its value in reducing non-performing assets.

→ Similarly, in our project, Logistic Regression also outperformed other models (~95%, cutoff 0.5), reinforcing its effectiveness in loan approval prediction.

Orji (2022) found Random Forest most accurate (95.55%) and Logistic Regression lowest (80%) in loan approval prediction. In contrast, our project ranked Logistic Regression highest (~95%) and Random Forest second (~91%)

→ Model performance can vary by dataset and feature selection.

Implications for Theory: Credit Risk & Lending Behavior






Creditworthiness is multidimensional:

Variables like CreditScore and AnnualIncome alone don't dictate outcomes. Lenders consider combinations of factors, including behavior-based metrics like DebtToIncomeRatio.

Support for decision-theoretic frameworks:

The model's accuracy and interpretability validate data-driven decision-making approaches commonly used in financial services to improve consistency in lending decisions.

Implications - Practical

-  **Faster Loan Processing**
-  **Smarter Risk Management** Assess applicant credit risk
-  **Fair & Consistent Decisions** Reduce human bias
-  **Targeted Outreach** Market to low-risk applicants
-  **Policy Insights** Tune strategies based on key features

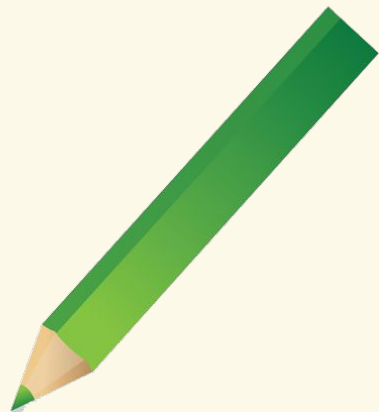
LOAN



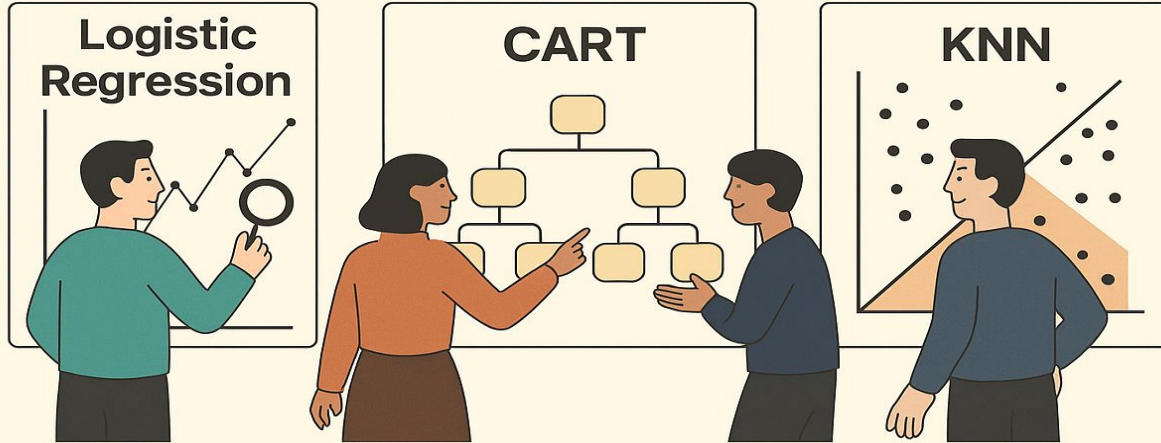
APPROVED



REJECTED



What This Means For Policy



- **Logistic Regression** gave balanced, stable results
- **CART** delivered better sensitivity
- **KNN** showed high specificity but low recall.
- Playing with probability cutoffs (0.3, 0.1) allowed us to simulate strict vs. lenient lending strategies.
- A data-driven approach supports consistent and fair lending decisions.

Limitation

Unbalance Data:

Rejected (0)	Approved(1)
6633	1817

Results	Accuracy	Specificity	Sensitivity
Cutoff 0.5	84.18%	97.24%	39.75%
Cutoff 0.3	84.0%	98.9%	33.33%
Cutoff 0.1	78.8%	88.1%	47.04%

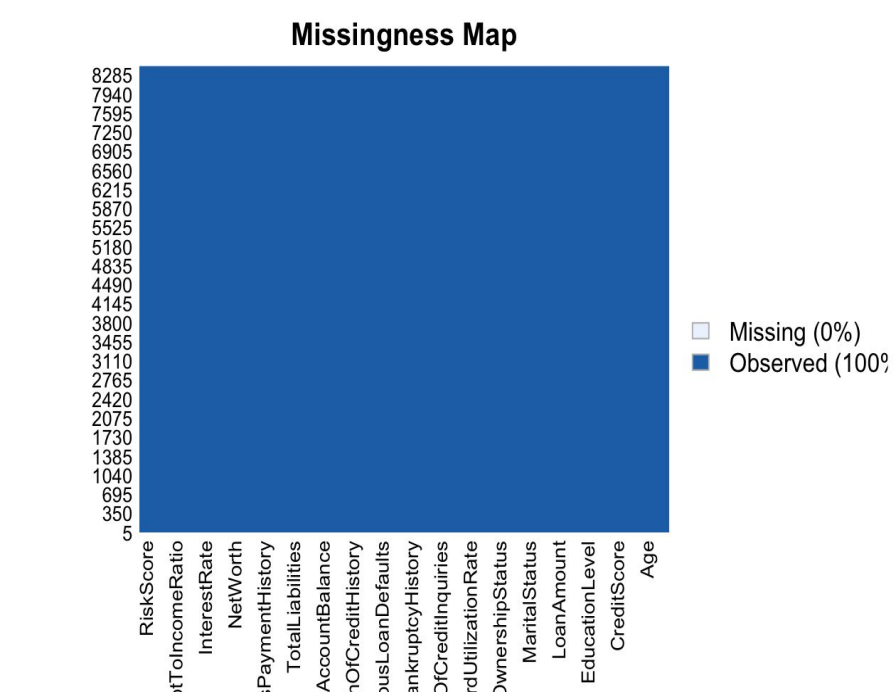
Limitation

No Missing Data

Fully Correct Data

Removed Outlier

Original	Reprocessing
20000	8450



FUTURE RESEARCH



- Ensemble methods like Random forest etc. can be explored.
- Use of Resampling methods.
- Optimizing cutoffs based on real-world cost-benefit.
- Combining two models.
- Including behavioral or time-based features.
- Evaluating model's fairness across different borrower groups.

SOURCES

M. A. Sheikh, A. K. Goel and T. Kumar, *An approach for prediction of loan approval using machine learning algorithm*. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)* (pp. 746–751). IEEE.

<https://doi.org/10.1109/ICESC48915.2020.9155614>

Haque, F. M. A., & Hassan, M. M. (2024). *Bank Loan Prediction Using Machine Learning Techniques*. American Journal of Industrial and Business Management, 14(12), 1421–1432. <https://doi.org/10.4236/ajibm.2024.1412085>

Nureni, A. A., & Adekola, O. E. (2022). Loan Approval Prediction Based on Machine Learning Approach. *Fudma Journal of Sciences*, 6, 41-50.

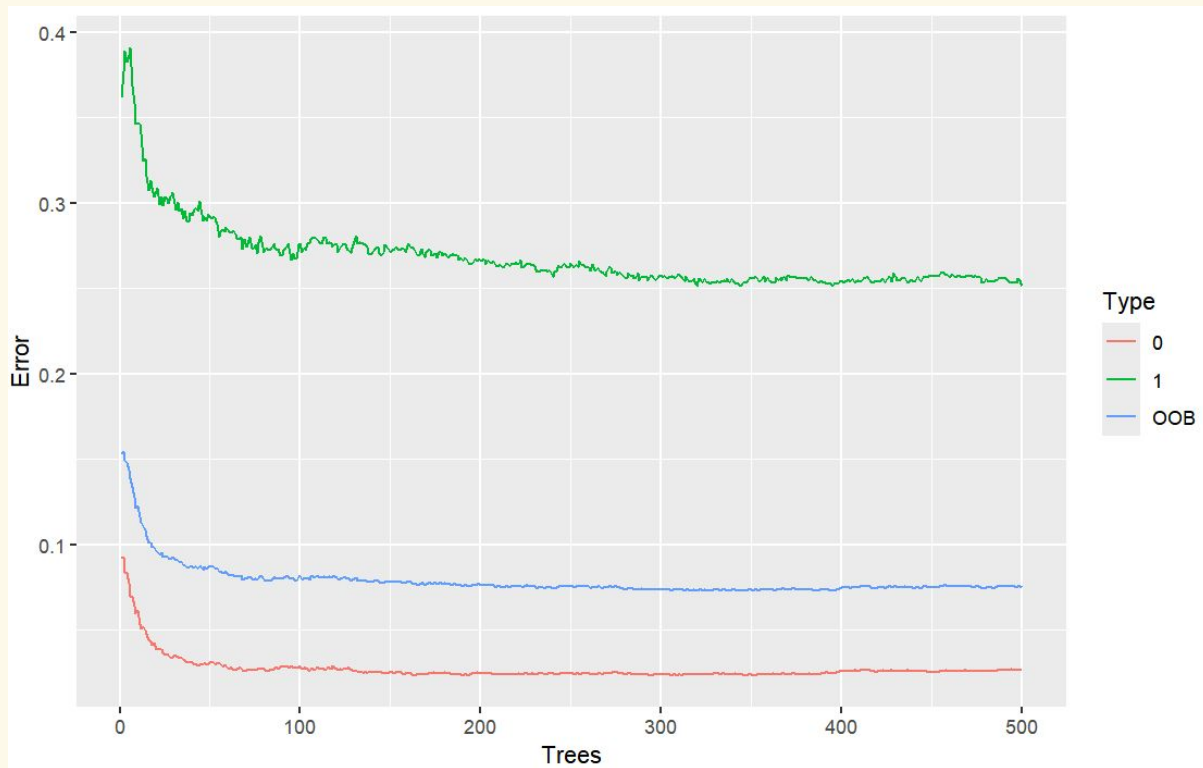
<https://doi.org/10.33003/fjs-2022-0603-830>

Orji, U. E., Ugwuishiwu, C. H., Nguemaleu, J. C. N., & Ugwuanyi, P. N. (2022). Machine Learning Models for Predicting Bank Loan Eligibility. In *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)* (pp. 1-5). IEEE.

<https://doi.org/10.1109/nigercon54645.2022.9803172>

THANK YOU.
Questions?

OOB Error Plot



OOB scenario for mtry 1:15

```
> oob.values
```

```
[1] 0.21318681 0.10329670 0.08453085 0.07726120 0.07523246 0.07421809 0.07404903 0.07218935 0.07218935  
[10] 0.07117498 0.07371090 0.07117498 0.07303466 0.07202029 0.07218935
```

Default vs Tuning

		Default Setting (mtry= 6, number of tree = 500)	Tuning Model (mtry = 10, number of tree = 1000)
Cutoff= 0.5	Accuracy Sensitivity Specificity	91.6% 74.4% 96.4%	91.6% 75.1% 96.1%
Cutoff= 0.3	Accuracy Sensitivity Specificity	89.7% 89.6% 89.7%	89.9% 89.5% 90.0%
Cutoff= 0.1	Accuracy Sensitivity Specificity	76.6% 98.9% 70.5%	78.6% 99.1% 72.9%