CREDIT STANDING

PREDICTIVE ANALYSIS

INTRODUCTION

BUILDING A PREDICTIVE MODEL THAT CAN ASSESS THE CREDIT WORTHINESS OF POTENTIAL FUTURE CUSTOMERS OF A FINANCIAL INSTITUTION

CONTENT INDEX

GOAL

- THE FINANCIAL INSTITUTION CAN DEVELOP A MACHINE LEARNING MODEL THAT USES HISTORICAL LOAN DATA TO PREDICT THE LIKELIHOOD OF LOAN DEFAULT.

- THE MODEL CAN BE TRAINED ON A RANGE OF FACTORS SUCH AS THE APPLICANT'S CREDIT HISTORY, INCOME, EMPLOYMENT STATUS, LOAN AMOUNT, LOAN DURATION, ETC. THE MODEL CAN USE DIFFERENT ALGORITHMS SUCH AS LOGISTIC REGRESSION, DECISION TREES, OR NEURAL NETWORKS TO LEARN THE PATTERNS AND RELATIONSHIPS BETWEEN THESE VARIABLES AND THE LIKELIHOOD OF LOAN DEFAULT.

USE CASE

- ONCE THE MODEL IS TRAINED AND VALIDATED, IT CAN BE USED TO SCORE NEW LOAN APPLICATIONS AUTOMATICALLY.

- THE MODEL CAN GENERATE A CREDIT SCORE FOR EACH APPLICANT, INDICATING THEIR CREDIT WORTHINESS AND THE RISK OF DEFAULT.

- BASED ON THESE SCORES, THE FINANCIAL INSTITUTION CAN DECIDE WHETHER TO APPROVE OR REJECT THE LOAN APPLICATION, OR ADJUST THE LOAN TERMS SUCH AS INTEREST RATE OR LOAN DURATION.

ABOUT DATA

THE DATASET CONTAINS INFORMATION ON CUSTOMERS' FINANCIAL STANDING, EMPLOYMENT, PERSONAL STATUS, AND DEMOGRAPHIC INFORMATION.

THE DATASET CONTAINS INFORMATION ON 807 PAST LOAN CUSTOMER CASES, WITH 14 ATTRIBUTES FOR EACH CASE.

DATA CLEANING

MISSING VALUES

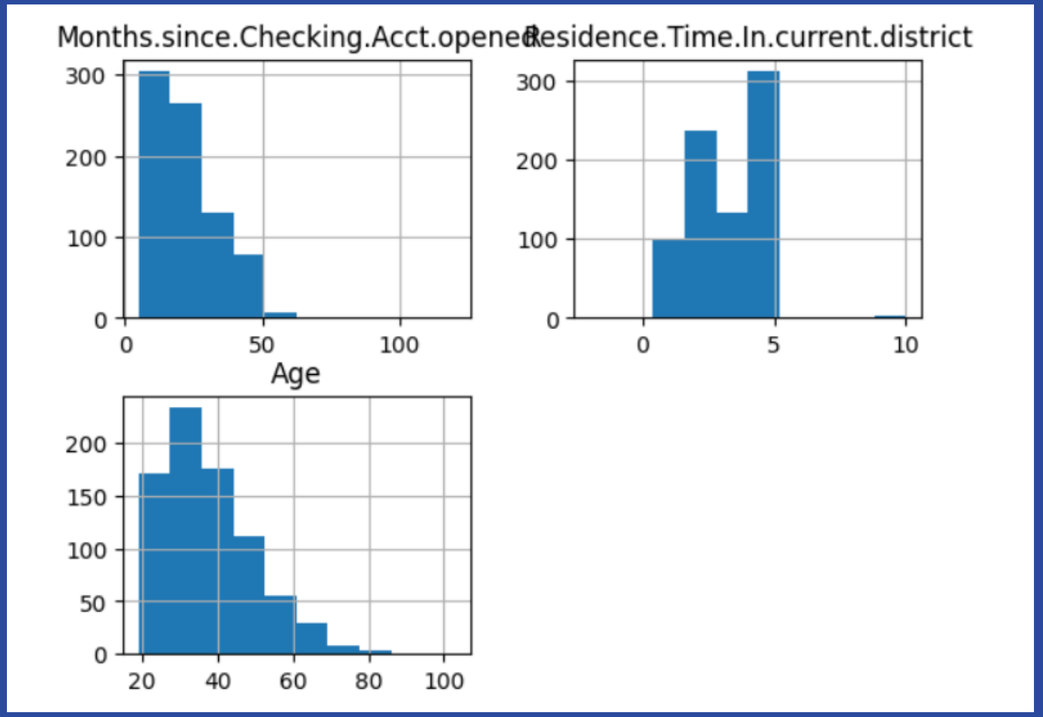
THE NUMBER OF MISSING VALUES IS VERY SMALL AND THERE IS NO MEANINGFUL PATTERN TO THEIR OCCURRENCE THEREFORE, DELETING MISSING VALUES CAN BE A REASONABLE APPROACH.

SKEWNESS

THE SKEWNESS VALUES SUGGEST THAT THE "MONTHS.SINCE.CHECKING.ACCT.OPENED" COLUMN HAS THE MOST EXTREME VALUES ON THE RIGHT SIDE OF THE DISTRIBUTION,

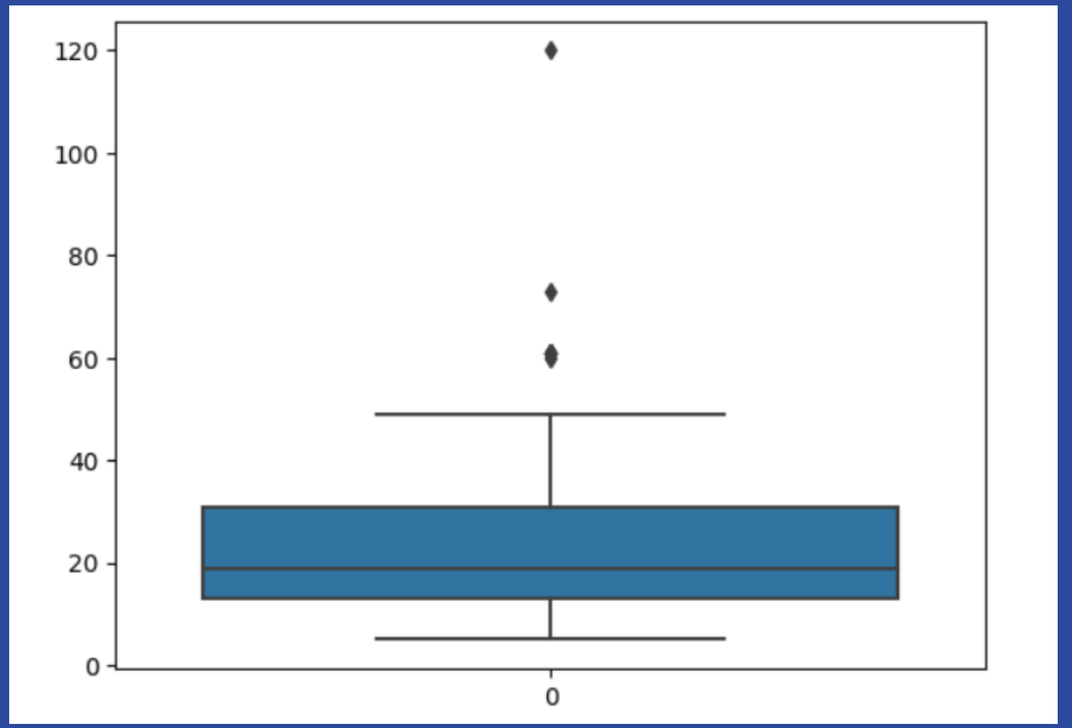
"RESIDENCE.TIME.IN.CURRENT.DISTRICT" COLUMN HAS A NEARLY SYMMETRICAL DISTRIBUTION,

THE "AGE" COLUMN IS MODERATELY SKEWED.



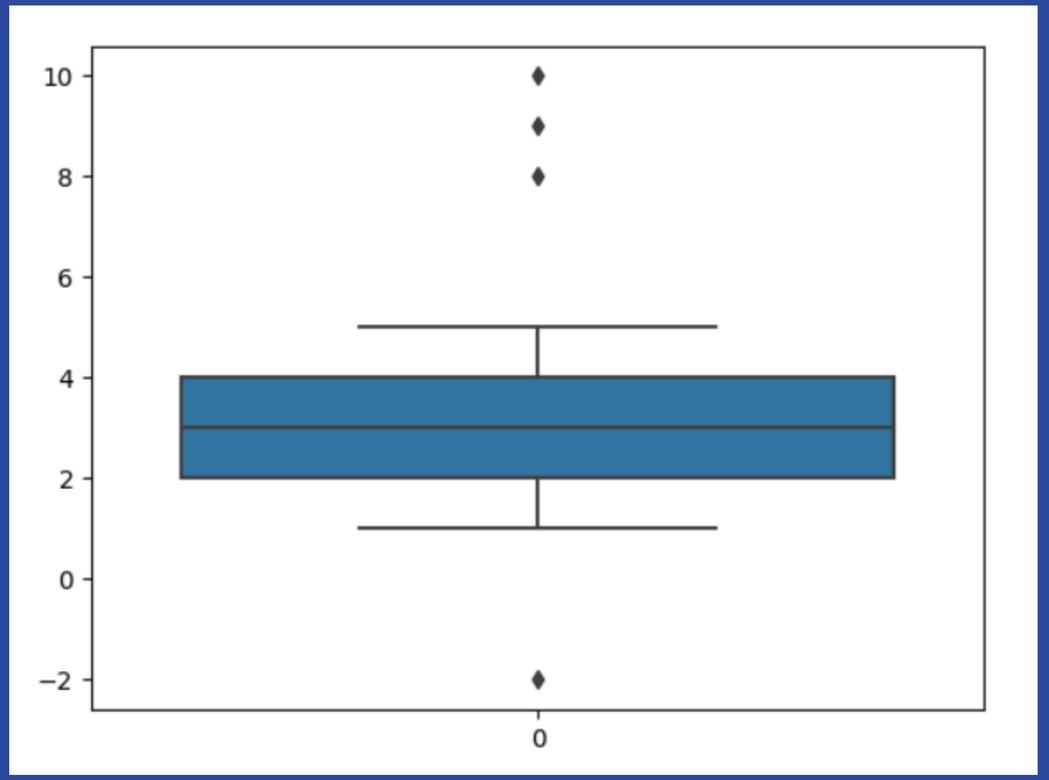
OUTLIERS

THE VARIABLE "MONTHS.SINCE.CHECKING.ACCT.OPENED", THERE IS 1 OUTLIER WITH A VALUE OF 0.475885.



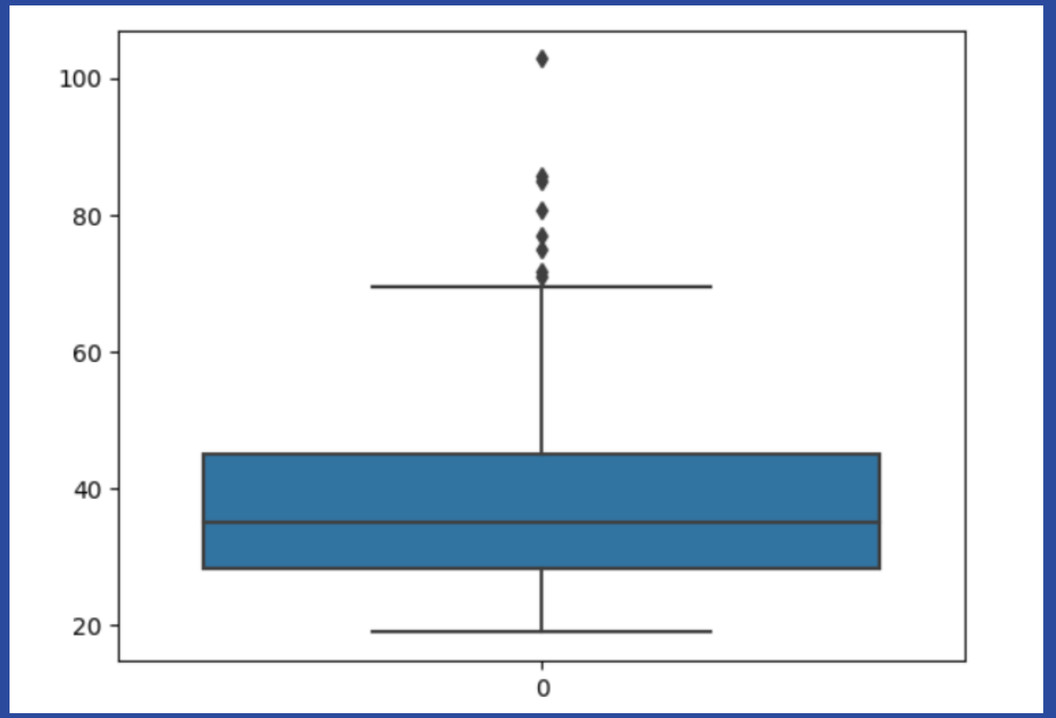
OUTLIERS

THE VARIABLE "RESIDENCE.TIME.IN.CURRENT.DISTRICT", THERE ARE 4 OUTLIERS WITH VALUES OF -2, 10, 9, AND 8.



OUTLIERS

THE VARIABLE "AGE", THERE IS 1 OUTLIER WITH A VALUE OF 4.63264907.



DATA TRANSFORMATION

1

FACTORIZED THE CATEGORICAL DATA

done to make the variable easier to work with in statistical analysis or machine learning models.

2

DATA ENCODING

each category or level of a categorical variable is assigned a unique numerical value

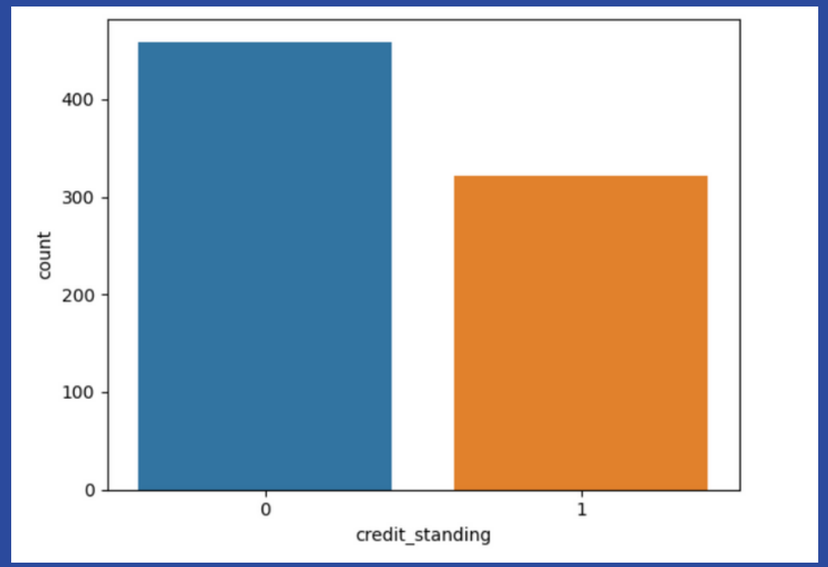
CLIENTS

OUR THOROUGH SERVICES HAVE WON THE AMAZEMENT AND LOYALTY FROM THE CLIENTS OVER THE DECADES

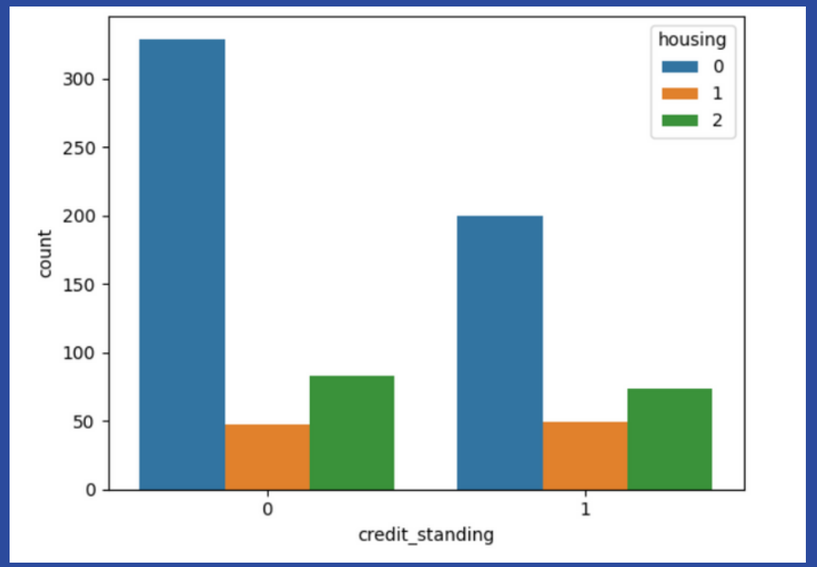
- Borcelle
- Ingoude Company
- Keithston and Partners
- Larana, Inc.
- Paucek and Lage
- Warner & Spencer

EXPLORATORY DATA ANALYSIS

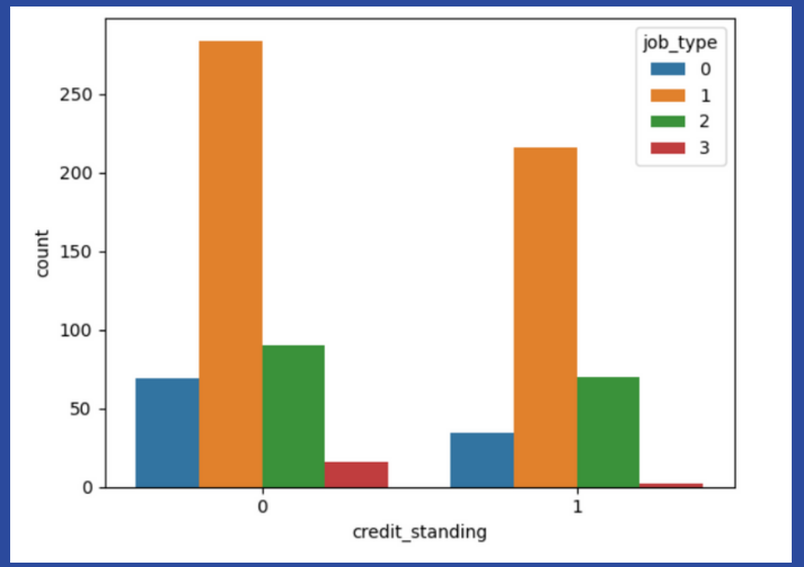
STANDING DISTRIBUTION



HOUSING AS HUE

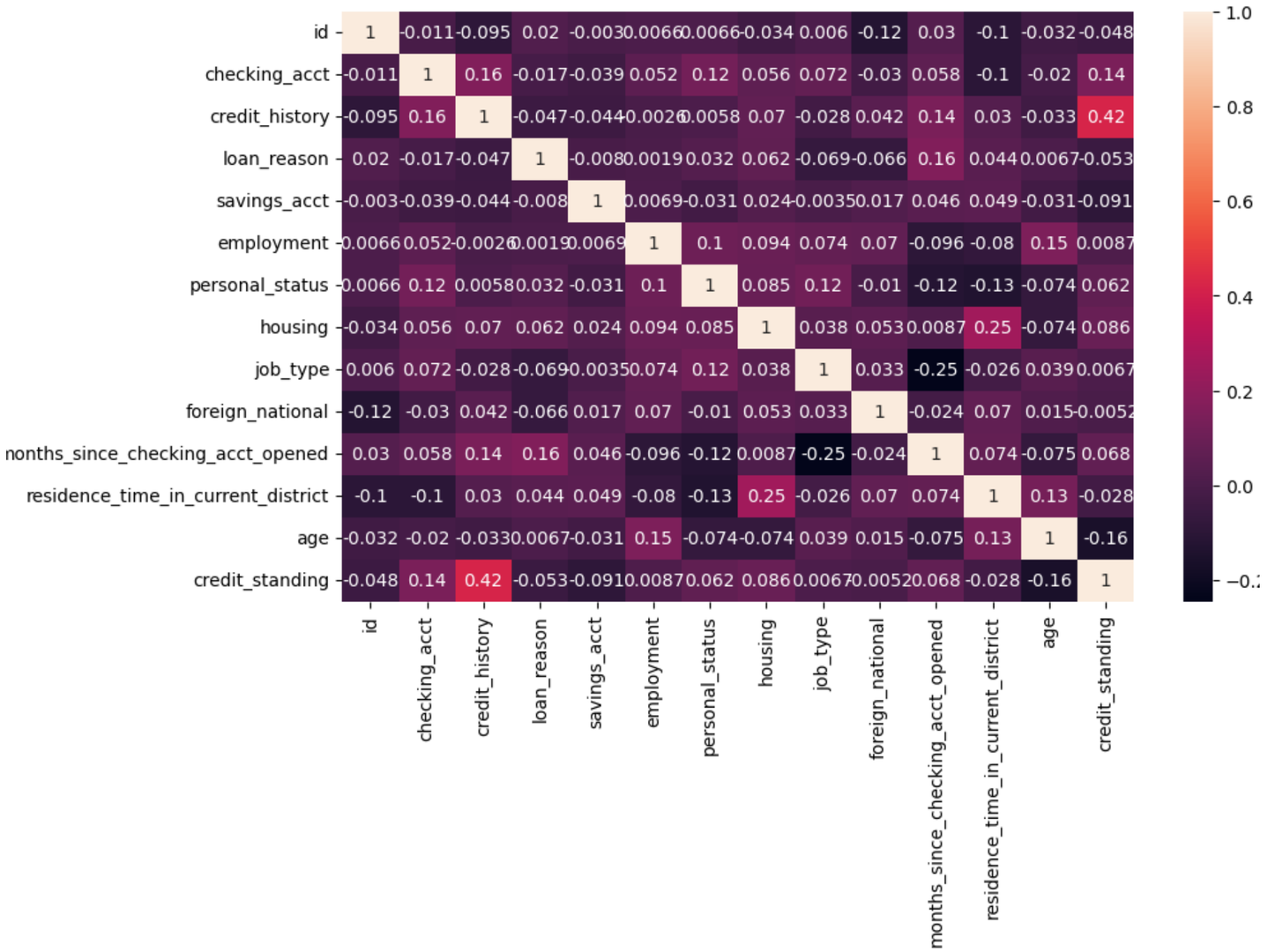


JOB TYPE AS HUE



CORRELATIONS

THERE IS NO STRONG NEGATIVE OR POSITIVE CORRELATION BETWEEN THESE VARIABLES AND HENCE NO LINEAR RELATIONSHIP.



MODELS TRAINED

1 **MODEL 1: LOGISTIC REGRESSION**

Logistic Regression Accuracy: 66.35 %

2 **MODEL 2: KNN**

K-Nearest Neighbors Accuracy: 65.40 %

3 **MODEL 3: DECISION TREE**

Decision Tree Accuracy: 75.50 %

MODELS TRAINED

1 **MODEL 4: RANDOM FOREST**

Random Forest Accuracy: 76.79 %

2 **MODEL 5: SUPPORT VECTOR MACHINE**

Support Vector Machine Accuracy: 70.68 %

3 **MODEL 6: NAIVE BAYES**

Naive Bayes Accuracy: 64.42 %

PROPOSED MODELING TECHNIQUE

DECISION TREE AND RANDOM FOREST ARE PROVIDING ALMOST SIMILAR ACCURACY BUT THE RANDOM FOREST WOULD BE THE BETTER CHOICE FOR FOLLOWING REASONS:

- Random Forest is a versatile algorithm that can handle both classification and regression tasks.
- It can also work well with both numerical and categorical data.
- Random Forest is an ensemble learning method that combines multiple decision trees to make predictions.
- By combining multiple trees, it reduces the risk of overfitting, which can occur when a model learns the training data too well and performs poorly on new data.
- Random Forest can handle missing data well. It can make use of available data to predict missing values and does not require imputation of missing data.
- Random Forest is less sensitive to outliers compared to other models like linear regression
- Random Forest provides a measure of feature importance, which can be useful in understanding the most important features
- Random Forest can handle large datasets with many features efficiently.

BENEFITS

1. **IMPROVED ACCURACY:** MACHINE LEARNING MODELS CAN LEARN COMPLEX PATTERNS AND RELATIONSHIPS BETWEEN VARIABLES THAT ARE DIFFICULT TO CAPTURE WITH TRADITIONAL STATISTICAL METHODS. THIS CAN IMPROVE THE ACCURACY OF CREDIT SCORING AND REDUCE THE RISK OF DEFAULT.
2. **FASTER PROCESSING:** AUTOMATED CREDIT SCORING CAN SPEED UP THE LOAN APPROVAL PROCESS AND REDUCE THE TIME AND COST OF MANUAL ASSESSMENT.
3. **CONSISTENCY:** MACHINE LEARNING MODELS CAN APPLY CONSISTENT AND OBJECTIVE CRITERIA TO ALL LOAN APPLICATIONS, ELIMINATING SUBJECTIVE BIAS OR ERRORS THAT MAY OCCUR IN MANUAL ASSESSMENT.
4. **BETTER PORTFOLIO MANAGEMENT:** BY ACCURATELY PREDICTING CREDIT WORTHINESS, THE FINANCIAL INSTITUTION CAN IMPROVE ITS LOAN PORTFOLIO MANAGEMENT AND REDUCE THE RISK OF BAD LOANS.
5. **INCREASED CUSTOMER SATISFACTION:** FASTER AND MORE ACCURATE LOAN APPROVAL CAN IMPROVE CUSTOMER SATISFACTION AND ATTRACT MORE BUSINESS.

CONCLUSION

IN CONCLUSION, IMPLEMENTING A MACHINE LEARNING MODEL FOR CREDIT SCORING CAN PROVIDE SIGNIFICANT BENEFITS TO FINANCIAL INSTITUTIONS BY IMPROVING THE ACCURACY AND SPEED OF LOAN PROCESSING, REDUCING THE RISK OF DEFAULT, AND INCREASING CUSTOMER SATISFACTION.