# MACHINE LEARNING PROJECT

## Loan

**INTRODUCTION**

*PROBLEM*

Whenever a person visits any financial institution to get a loan, there is a lot of paper work involved, and it is a time-consuming process as well. But even after all these, there is still ambiguity about whether the loan will be approved or not. In case the loan isn't approved, a lot of time and resources are wasted in the whole process.

*DATASET AND VARIABLES*

The dataset is named as 'LoanDefaulters' and contains past records of people who applied for loan with all their information and whether their loan was approved or not approved.

The dataset has the following variables:

| S. No. | Variable Name | Description |
|---|---|---|
| 1 | ID | Loan Borrower's ID |
| 2 | Gender | Gender |
| 3 | approv_in_adv | Approved in Advance (non pre or pre) |
| 4 | loan_type | Loan Type (Type 1 or Type 2 or Type 3) |
| 5 | loan_purpose | Purpose of Loan (p1 / p2 / p3 / p4) |
| 6 | Credit_Worthiness | Type of Credit worthiness (I1 or I2) |
| 7 | open_credit | Open Credit or not (opc / nopc) |
| 8 | business_or_commercial | Business loan or commercial loan |
| 9 | loan_amount | Loan Amount |
| 10 | rate_of_interest | Interest Rate |
| 11 | property_value | Property Value |
| 12 | income | Borrower's Income |
| 13 | credit_type | Credit Type |

| | | |
|---|---|---|
| 14 | Credit_Score | Credit Score |
| 15 | age | Borrower's Age |
| 16 | LTV | Loan to Value Ratio |
| 17 | Region | Borrower's Region |
| 18 | Security_Type | Security Type |
| 19 | Status | Loan Status (0: Not approved; 1: Approved) |

These variables can be further categorized into the following types:

**Discrete, Categorical, Continuous**

| S. No. | Variable Name | Description |
|---|---|---|
| 1 | ID | Loan Borrower’s ID |
| 2 | Gender | Gender |
| 3 | approv_in_adv | Approved in Advance (non pre or pre) |
| 4 | loan_type | Loan Type (Type 1 or Type 2 or Type 3) |
| 5 | loan_purpose | Purpose of Loan (p1 / p2 / p3 / p4) |
| 6 | Credit_Worthiness | Type of Credit worthiness (I1 or I2) |
| 7 | open_credit | Open Credit or not (opc / nopc) |
| 8 | business_or_commercial | Business loan or commercial loan |
| 9 | loan_amount | Loan Amount |
| 10 | rate_of_interest | Interest Rate |
| 11 | property_value | Property Value |
| 12 | Income | Borrower's Income |
| 13 | credit_type | Credit Type |
| 14 | Credit_Score | Credit Score |
| 15 | Age | Borrower's Age |
| 16 | LTV | Loan to Value Ratio |
| 17 | Region | Borrower's Region |
| 18 | Security_Type | Security Type |
| 19 | Status | Loan Status (0:Not approved; 1: Approved) |

## OBJECTIVE

With the help of this dataset, we aim to create a predictive model that can predict dependent variable; that is status of the loan, with the help of all other independent variables. And using our prediction model a person can know if he should apply for the loan based on his chances of approval. So, that he can save his resources and time if there are less chances of loan approval.

## METHODOLOGY

### *DATASET CLEANING*

There were no duplicate rows in the entire dataset.

The maximum number of missing values were in ‘rate_of_interest’ variable, i.e., 718 values. The whole dataset has 2892 observations; and removing 718 observations, will make the dataset smaller and hence we will only use imputation of missing values and not removal of them.

### *IDENTIFICATION OF VARIABLES*

Status (approved or not approved) will be the dependent variable. Out of the remaining variables, ID is a discrete variable and signifies customer ID which won’t be used as an independent variable. Apart from ID, Security_Type has only 1 level of factor and hence won’t be used as an independent variable.

The remining variables that are; Gender, approv_in_adv, loan_type, loan_purpose, Credit_Worthiness, open_credit, business_or_commercial, loan_amount, rate_of_interest, property_value, income, credit_type, Credit_Score, age, LTV and Region; can be used as independent variables.

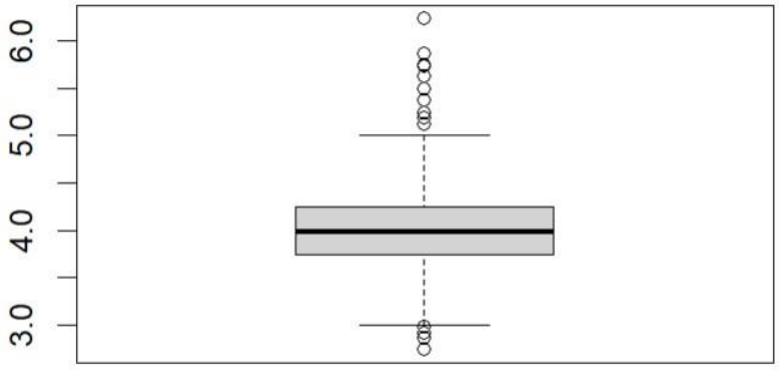
Out of there independent variable only 6 are continuous (loan_amount, rate_of_interest, property_value, income, Credit_Score and LTV) and needs to be tested for normality using skewness and kurtosis.

```
> skewness(L1[c(9,10,11,12,14,16)])
     loan_amount rate_of_interest   property_value           income
     1.532701796      0.353708200      4.169592057     12.255668646
    Credit_Score              LTV
    -0.005807775     -0.827689483
> kurtosis(L1[c(9,10,11,12,14,16)])
     loan_amount rate_of_interest   property_value           income
        9.290111         3.918559        41.252515       265.651109
    Credit_Score              LTV
        1.829448         3.978403
```

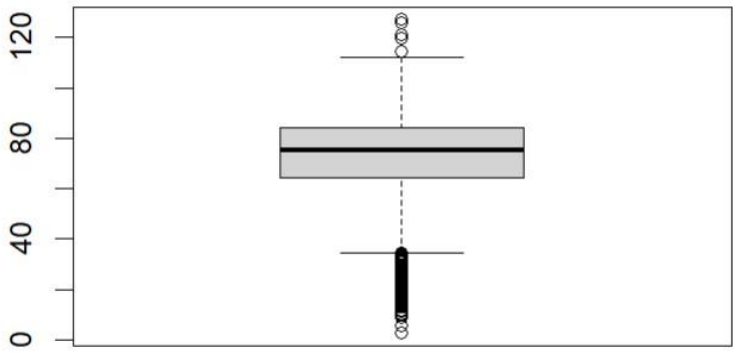
Out of these 6 continuous variables, 3 are normal (rate_of_interest, Credit_Score and LTV) while the remaining three (loan_amount, property_value and income) are not normal and must be treated for the same.

Out of the 3 continuous normal variables the following two have moderate outliers:

rate_of_interest (109 outliers)




LTV (111 outliers)




*POSSIBLE TREATMENT FOR THE IDENTIFIED ISSUES*

Loan Amount (32 outliers present) – sqrt / log (17 / 17 outliers still present), imputation of outliers (0 outliers), removal of outliers (0 outliers)

Property value (121 outliers present) – log (58 outliers still present); removal of outliers – normality achieved (25 outliers still present); imputation of outliers – normality achieved (54 outliers still present)

Income (152 outliers present) – Direct transformation not possible; removal of outliers – normality achieved (47 outliers still present); imputation of outliers – normality achieved (72 outliers still present)

** Income includes outlier of loan_amount as well; and hence removing its outlier will treat loan_amount as well.

As for outliers in rate_of_interest and LTV; they are only moderate and are less than 4 percent of total observation, so we will make two categories of models; first in which outliers aren't treated and second in which outlier values are treated using imputation.

*POSSIBLE DATASETS WITH DIFFERENT TREATMENT*

| Dataset No. | Loan amount | Property value | Income | LTV | Rate of interest |
|---|---|---|---|---|---|
| 1 | - | - | - | - | - |
| 2 | - | log | removal of outliers | | |
| 3 | sqrt | log | imputation | | |
| 4 | - | removal of outliers | removal of outliers | | |
| 5 | sqrt | removal of outliers | imputation | | |
| 6 | - | imputation | removal of outliers | | |
| 7 | sqrt | imputation | imputation | | |
| 8 | log | log | imputation | | |

| | | | | | |
|---|---|---|---|---|---|
| 9 | log | removal of outliers | imputation | | |
| 10 | log | imputation | imputation | | |
| 11 | imputation | log | removal of outliers | | |
| 12 | imputation | log | imputation | | |
| 13 | imputation | removal of outliers | removal of outliers | | |
| 14 | imputation | removal of outliers | imputation | | |
| 15 | imputation | imputation | removal of outliers | | |
| 16 | imputation | imputation | imputation | | |
| 17 | removal | log | removal of outliers | | |
| 18 | removal | log | imputation | | |
| 19 | removal | removal of outliers | removal of outliers | | |
| 20 | removal | removal of outliers | imputation | | |
| 21 | removal | imputation | removal of outliers | | |
| 22 | removal | imputation | imputation | | |
| 23 | - | log | removal of outliers | imputation | imputation |
| 24 | sqrt | log | imputation | imputation | imputation |
| 25 | - | removal of outliers | removal of outliers | imputation | imputation |
| 26 | sqrt | removal of outliers | imputation | imputation | imputation |
| 27 | - | imputation | removal of outliers | imputation | imputation |
| 28 | sqrt | imputation | imputation | imputation | imputation |
| 29 | log | log | imputation | imputation | imputation |
| 30 | log | removal of outliers | imputation | imputation | imputation |
| 31 | log | imputation | imputation | imputation | imputation |
| 32 | imputation | log | removal of outliers | imputation | imputation |
| 33 | imputation | log | imputation | imputation | imputation |
| 34 | imputation | removal of outliers | removal of outliers | imputation | imputation |
| 35 | imputation | removal of outliers | imputation | imputation | imputation |
| 36 | imputation | imputation | removal of outliers | imputation | imputation |
| 37 | imputation | imputation | imputation | imputation | imputation |
| 38 | removal | log | removal of outliers | imputation | imputation |
| 39 | removal | log | imputation | imputation | imputation |
| 40 | removal | removal of outliers | removal of outliers | imputation | imputation |
| 41 | removal | removal of outliers | imputation | imputation | imputation |
| 42 | removal | imputation | removal of outliers | imputation | imputation |
| 43 | removal | imputation | imputation | imputation | imputation |

Dataset 1 is basic dataset with only imputation of missing values which can be used for decision trees and random forest only as they don't require normality assumption.

All other datasets are normal and don't have severe outliers and hence can be used for all methods.

*POSSIBLE METHODS FOR PREDICTIVE MODELS*

Status is dependent variables and is categorical (binomial); so logistic regression, naïve bayes, decision tree (Gini Index), decision tree (Information Gain) and random forest methods can be used.

*ASSUMPTION FOR THOSE MODELS*

Logistic regression:

Continuous IVs are normal

No severe outliers

No multicollinearity between IVs

Naïve Bayes:

Continuous IVs are normal

No severe outliers

No correlation between IVs (Since, no correlation is practically not possible. We only considered moderate and high correlation for this assumption)

Rest, decision trees and random forest don't have any assumption and can be used on any dataset.

*MODELS POSSIBLE FOR OUR DATASETS*

| **Logistic Regression** | | **Naïve Bayes** | | **Decision Tree** | | **Random Forest** | **Total** |
|---|---|---|---|---|---|---|---|
| *All variables* | *Significant variables* | *Loan amount* | *Property value* | *Gini Index* | *Information Gain* | | |
| 42 models | 42 models | 42 models | 42 models | 43 models | 43 models | 43 models | **297 models** |

Logistic Regression (84 models):

Out of 43 datasets; 42 are normally distributed and don't have any issue of multicollinearity and can be used for logistic regression. 84 models can be created using these 42 datasets. Initial 42 models with all the IVs in the dataset and the later with only significant IVs of those datasets.

Naïve Bayes (84 models):

Out of 43 datasets; 42 are normally distributed. But in these 42 datasets 'property value' and 'loan amount' are moderately correlated (r>0.7) and hence can't be used together and hence two models are created using each dataset. One model without property value and the other without loan amount. And hence total 84 models can be created.

Decision tree (86 models):

All 43 datasets can be used for decision tree. Decision tree can be made by gini index and information gain and hence 2 models can be created using each dataset. A total of 86 models can be created.

Random forest (43 models):

All 43 datasets can be used for random forest. A total of 43 models can be created.

**DATA ANALYSIS**

We created all the possible models. Their accuracy, sensitivity and specificity are as following:

*Logistic Regression:*

(With all variables):

| Dataset | Loan amount | Property value | Income | LTV | Rate of interest | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|---|---|---|---|
| 2 | - | log | removal of outliers | | | 85.74 | 44.53 | 99.51 |
| 3 | sqrt | log | imputation | | | 85.12 | 42.36 | 99.31 |
| 4 | - | removal of outliers | removal of outliers | | | 87.33 | 51.49 | 99.49 |
| 5 | sqrt | removal of outliers | imputation | | | 85.74 | 44.29 | 99.76 |
| 6 | - | imputation | removal of outliers | | | 85.92 | 44.53 | 99.76 |
| 7 | sqrt | imputation | imputation | | | 85.12 | 42.36 | 99.31 |
| 8 | log | log | imputation | | | 85.47 | 43.06 | 99.54 |
| 9 | log | removal of outliers | imputation | | | 85.92 | 45 | 99.76 |
| 10 | log | imputation | imputation | | | 85.29 | 42.36 | 99.54 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 11 | imputation | log | removal of outliers | | | 85.37 | 43.8 | 99.27 |
| 12 | imputation | log | imputation | | | 85.29 | 42.36 | 99.54 |
| 13 | imputation | removal of outliers | removal of outliers | | | 86.96 | 51.49 | 98.99 |
| 14 | imputation | removal of outliers | imputation | | | 85.38 | 42.57 | 99.52 |
| 15 | imputation | imputation | removal of outliers | | | 85.92 | 44.53 | 99.76 |
| 16 | imputation | imputation | imputation | | | 85.29 | 42.36 | 99.54 |
| 17 | removal | log | removal of outliers | | | 84.56 | 37.78 | 100 |
| 18 | removal | log | imputation | | | 84.62 | 39.72 | 99.3 |
| 19 | removal | removal of outliers | removal of outliers | | | 87.12 | 50.38 | 99.49 |
| 20 | removal | removal of outliers | imputation | | | 86.41 | 47.83 | 99.28 |
| 21 | removal | imputation | removal of outliers | | | 84.56 | 37.77 | 100 |
| 22 | removal | imputation | imputation | | | 84.97 | 40.42 | 99.53 |
| 23 | - | log | removal of outliers | imputation | imputation | 85.37 | 43.8 | 99.27 |
| 24 | sqrt | log | imputation | imputation | imputation | 85.29 | 43.06 | 99.31 |
| 25 | - | removal of outliers | removal of outliers | imputation | imputation | 87.15 | 50.75 | 99.49 |
| 26 | sqrt | removal of outliers | imputation | imputation | imputation | 85.02 | 42.14 | 99.52 |
| 27 | - | imputation | removal of outliers | imputation | imputation | 85.92 | 43.8 | 100 |
| 28 | sqrt | imputation | imputation | imputation | imputation | 85.12 | 42.36 | 99.31 |
| 29 | log | log | imputation | imputation | imputation | 85.47 | 43.06 | 99.54 |
| 30 | log | removal of outliers | imputation | imputation | imputation | 85.02 | 42.86 | 99.28 |
| 31 | log | imputation | imputation | imputation | imputation | 85.29 | 42.36 | 99.54 |
| 32 | imputation | log | removal of outliers | imputation | imputation | 85.56 | 43.07 | 99.76 |
| 33 | imputation | log | imputation | imputation | imputation | 85.47 | 43.06 | 99.54 |
| 34 | imputation | removal of outliers | removal of outliers | imputation | imputation | 87.15 | 50.75 | 99.49 |
| 35 | imputation | removal of outliers | imputation | imputation | imputation | 85.02 | 42.14 | 99.52 |
| 36 | imputation | imputation | removal of outliers | imputation | imputation | 85.74 | 43.07 | 100 |
| 37 | imputation | imputation | imputation | imputation | imputation | 85.29 | 42.36 | 99.54 |
| 38 | removal | log | removal of outliers | imputation | imputation | 84.01 | 37.77 | 99.26 |
| 39 | removal | log | imputation | imputation | imputation | 84.44 | 39 | 99.3 |
| 40 | removal | removal of outliers | removal of outliers | imputation | imputation | 87.12 | 50.38 | 99.49 |
| 41 | removal | removal of outliers | imputation | imputation | imputation | 86.59 | 47.83 | 99.52 |
| 42 | removal | imputation | removal of outliers | imputation | imputation | 83.82 | 37.03 | 99.26 |
| 43 | removal | imputation | imputation | imputation | imputation | 84.62 | 39 | 99.53 |

(With significant variables):

| Dataset | Loan amount | Property value | Income | LTV | Rate of interes | Sig Accuracy | Sig Sensitivity | Sig Specificity |
|---|---|---|---|---|---|---|---|---|
| 2 | - | log | removal of outliers | | | 75.32 | 2.18 | 99.75 |
| 3 | sqrt | log | imputation | | | 85.64 | 42.36 | 100 |
| 4 | - | removal of outliers | removal of outliers | | | 75.05 | 1.49 | 100 |
| 5 | sqrt | removal of outliers | imputation | | | 74.19 | 4.28 | 97.82 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 6 | - | imputation | removal of outliers | | | 75.32 | 2.18 | 99.75 |
| 7 | sqrt | imputation | imputation | | | 85.47 | 42.36 | 99.77 |
| 8 | log | log | imputation | | | 85.64 | 43.06 | 99.77 |
| 9 | log | removal of outliers | imputation | | | 74.19 | 4.28 | 97.82 |
| 10 | log | imputation | imputation | | | 75.09 | 2.08 | 99.3 |
| 11 | imputation | log | removal of outliers | | | 74.95 | 0 | 100 |
| 12 | imputation | log | imputation | | | 85.47 | 42.36 | 99.77 |
| 13 | imputation | removal of outliers | removal of outliers | | | 75.05 | 1.49 | 100 |
| 14 | imputation | removal of outliers | imputation | | | 74.19 | 4.28 | 97.82 |
| 15 | imputation | imputation | removal of outliers | | | 75.14 | 2.91 | 99.26 |
| 16 | imputation | imputation | imputation | | | 85.47 | 42.36 | 99.77 |
| 17 | removal | log | removal of outliers | | | 75.74 | 2.22 | 100 |
| 18 | removal | log | imputation | | | 84.97 | 39.71 | 99.76 |
| 19 | removal | removal of outliers | removal of outliers | | | 74.62 | 3 | 98.73 |
| 20 | removal | removal of outliers | imputation | | | 74.46 | 2.17 | 98.55 |
| 21 | removal | imputation | removal of outliers | | | 75.74 | 2.22 | 100 |
| 22 | removal | imputation | imputation | | | 84.97 | 39.71 | 99.76 |
| 23 | - | log | removal of outliers | imputation | imputation | 74.95 | 0 | 100 |
| 24 | sqrt | log | imputation | imputation | imputation | 85.29 | 43.06 | 99.31 |
| 25 | - | removal of outliers | removal of outliers | imputation | imputation | 74.67 | 0.76 | 99.74 |
| 26 | sqrt | removal of outliers | imputation | imputation | imputation | 74.19 | 2.14 | 98.55 |
| 27 | - | imputation | removal of outliers | imputation | imputation | 74.95 | 0 | 100 |
| 28 | sqrt | imputation | imputation | imputation | imputation | 85.29 | 42.36 | 99.54 |
| 29 | log | log | imputation | imputation | imputation | 85.47 | 42.36 | 99.77 |
| 30 | log | removal of outliers | imputation | imputation | imputation | 74.19 | 2.14 | 98.55 |
| 31 | log | imputation | imputation | imputation | imputation | 85.47 | 42.36 | 99.77 |
| 32 | imputation | log | removal of outliers | imputation | imputation | 74.95 | 0 | 100 |
| 33 | imputation | log | imputation | imputation | imputation | 85.47 | 43.06 | 99.54 |
| 34 | imputation | removal of outliers | removal of outliers | imputation | imputation | 74.86 | 0.74 | 100 |
| 35 | imputation | removal of outliers | imputation | imputation | imputation | 73.83 | 2.85 | 97.82 |
| 36 | imputation | imputation | removal of outliers | imputation | imputation | 75.14 | 0.72 | 100 |
| 37 | imputation | imputation | imputation | imputation | imputation | 85.47 | 42.36 | 99.77 |
| 38 | removal | log | removal of outliers | imputation | imputation | 75.37 | 0.74 | 100 |
| 39 | removal | log | imputation | imputation | imputation | 84.44 | 39 | 99.3 |
| 40 | removal | removal of outliers | removal of outliers | imputation | imputation | 74.05 | 1.5 | 98.48 |
| 41 | removal | removal of outliers | imputation | imputation | imputation | 74.46 | 1.44 | 98.79 |
| 42 | removal | imputation | removal of outliers | imputation | imputation | 75.37 | 0.74 | 100 |
| 43 | removal | imputation | imputation | imputation | imputation | 85.14 | 39.72 | 100 |

*Naïve Bayes:*

(With Loan Amount):

| Dataset | Loan amount | Property value | Income | LTV | Rate of interest | LA Accuracy | LA Sensitivity | LA Specificity |
|---|---|---|---|---|---|---|---|---|
| 2 | - | log | removal of outliers | | | 85.92 | 45.99 | 99.27 |
| 3 | sqrt | log | imputation | | | 85.99 | 45.14 | 99.54 |
| 4 | - | removal of outliers | removal of outliers | | | 86.96 | 52.24 | 98.73 |
| 5 | sqrt | removal of outliers | imputation | | | 85.02 | 40.71 | 100 |
| 6 | - | imputation | removal of outliers | | | 85.92 | 45.99 | 99.27 |
| 7 | sqrt | imputation | imputation | | | 85.99 | 45.14 | 99.54 |
| 8 | log | log | imputation | | | 85.99 | 45.14 | 99.54 |
| 9 | log | removal of outliers | imputation | | | 85.56 | 44.29 | 99.52 |
| 10 | log | imputation | imputation | | | 85.99 | 45.14 | 99.54 |
| 11 | imputation | log | removal of outliers | | | 85.92 | 45.99 | 99.27 |
| 12 | imputation | log | imputation | | | 85.99 | 45.14 | 99.54 |
| 13 | imputation | removal of outliers | removal of outliers | | | 86.96 | 52.24 | 98.73 |
| 14 | imputation | removal of outliers | imputation | | | 85.56 | 44.29 | 99.52 |
| 15 | imputation | imputation | removal of outliers | | | 85.92 | 45.99 | 99.27 |
| 16 | imputation | imputation | imputation | | | 85.81 | 43.06 | 100 |
| 17 | removal | log | removal of outliers | | | 84.19 | 40 | 98.77 |
| 18 | removal | log | imputation | | | 84.97 | 41.84 | 99.07 |
| 19 | removal | removal of outliers | removal of outliers | | | 86.93 | 54.14 | 97.97 |
| 20 | removal | removal of outliers | imputation | | | 87.86 | 52.9 | 99.52 |
| 21 | removal | imputation | removal of outliers | | | 84.19 | 40 | 98.77 |
| 22 | removal | imputation | imputation | | | 84.97 | 41.84 | 99.07 |
| 23 | - | log | removal of outliers | imputation | imputation | 85.92 | 45.99 | 99.27 |
| 24 | sqrt | log | imputation | imputation | imputation | 85.64 | 42.36 | 100 |
| 25 | - | removal of outliers | removal of outliers | imputation | imputation | 86.96 | 52.24 | 98.73 |
| 26 | sqrt | removal of outliers | imputation | imputation | imputation | 85.2 | 44.29 | 99.03 |
| 27 | - | imputation | removal of outliers | imputation | imputation | 85.92 | 45.99 | 99.27 |
| 28 | sqrt | imputation | imputation | imputation | imputation | 85.64 | 42.36 | 100 |
| 29 | log | log | imputation | imputation | imputation | 85.64 | 42.36 | 100 |
| 30 | log | removal of outliers | imputation | imputation | imputation | 85.2 | 44.29 | 99.03 |
| 31 | log | imputation | imputation | imputation | imputation | 85.64 | 42.36 | 100 |
| 32 | imputation | log | removal of outliers | imputation | imputation | 85.92 | 45.99 | 99.27 |
| 33 | imputation | log | imputation | imputation | imputation | 85.64 | 42.36 | 100 |
| 34 | imputation | removal of outliers | removal of outliers | imputation | imputation | 86.96 | 52.24 | 98.73 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 35 | imputation | removal of outliers | imputation | imputation | imputation | 85.2 | 44.29 | 99.03 |
| 36 | imputation | imputation | removal of outliers | imputation | imputation | 85.92 | 45.99 | 99.27 |
| 37 | imputation | imputation | imputation | imputation | imputation | 85.64 | 42.36 | 100 |
| 38 | removal | log | removal of outliers | imputation | imputation | 84.19 | 40 | 98.77 |
| 39 | removal | log | imputation | imputation | imputation | 84.97 | 42.55 | 98.84 |
| 40 | removal | removal of outliers | removal of outliers | imputation | imputation | 86.55 | 53.38 | 97.72 |
| 41 | removal | removal of outliers | imputation | imputation | imputation | 86.96 | 47.83 | 100 |
| 42 | removal | imputation | removal of outliers | imputation | imputation | 84.19 | 40 | 98.77 |
| 43 | removal | imputation | imputation | imputation | imputation | 84.97 | 42.55 | 98.84 |

(With property value):

| Dataset | Loan amount | Property value | Income | LTV | Rate of interes | PV Accuracy | PV Sensitivity | PV Specificity |
|---|---|---|---|---|---|---|---|---|
| 2 | - | log | removal of outliers | | | 85.92 | 45.99 | 99.27 |
| 3 | sqrt | log | imputation | | | 85.64 | 43.06 | 99.77 |
| 4 | - | removal of outliers | removal of outliers | | | 87.15 | 52.24 | 98.99 |
| 5 | sqrt | removal of outliers | imputation | | | 85.56 | 44.29 | 99.52 |
| 6 | - | imputation | removal of outliers | | | 86.11 | 45.99 | 99.51 |
| 7 | sqrt | imputation | imputation | | | 85.81 | 44.44 | 99.54 |
| 8 | log | log | imputation | | | 85.64 | 43.06 | 99.77 |
| 9 | log | removal of outliers | imputation | | | 85.56 | 44.29 | 99.52 |
| 10 | log | imputation | imputation | | | 85.64 | 43.06 | 99.77 |
| 11 | imputation | log | removal of outliers | | | 85.92 | 45.99 | 99.27 |
| 12 | imputation | log | imputation | | | 85.64 | 43.06 | 99.77 |
| 13 | imputation | removal of outliers | removal of outliers | | | 87.15 | 52.24 | 98.99 |
| 14 | imputation | removal of outliers | imputation | | | 85.02 | 40.71 | 100 |
| 15 | imputation | imputation | removal of outliers | | | 86.11 | 45.99 | 99.51 |
| 16 | imputation | imputation | imputation | | | 85.81 | 44.44 | 99.54 |
| 17 | removal | log | removal of outliers | | | 84.19 | 40 | 98.77 |
| 18 | removal | log | imputation | | | 84.97 | 41.84 | 99.07 |
| 19 | removal | removal of outliers | removal of outliers | | | 87.12 | 54.14 | 98.23 |
| 20 | removal | removal of outliers | imputation | | | 87.86 | 52.9 | 99.52 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 21 | removal | imputation | removal of outliers | | | 84.19 | 40 | 98.77 |
| 22 | removal | imputation | imputation | | | 84.97 | 41.84 | 99.07 |
| 23 | - | log | removal of outliers | imputation | imputation | 85.92 | 45.99 | 99.27 |
| 24 | sqrt | log | imputation | imputation | imputation | 85.99 | 45.14 | 99.54 |
| 25 | - | removal of outliers | removal of outliers | imputation | imputation | 86.96 | 52.24 | 98.73 |
| 26 | sqrt | removal of outliers | imputation | imputation | imputation | 84.84 | 40 | 100 |
| 27 | - | imputation | removal of outliers | imputation | imputation | 86.11 | 45.99 | 99.51 |
| 28 | sqrt | imputation | imputation | imputation | imputation | 85.81 | 44.44 | 99.54 |
| 29 | log | log | imputation | imputation | imputation | 85.99 | 45.14 | 99.54 |
| 30 | log | removal of outliers | imputation | imputation | imputation | 84.84 | 40 | 100 |
| 31 | log | imputation | imputation | imputation | imputation | 85.81 | 44.44 | 99.54 |
| 32 | imputation | log | removal of outliers | imputation | imputation | 85.92 | 45.99 | 99.27 |
| 33 | imputation | log | imputation | imputation | imputation | 85.99 | 45.14 | 99.54 |
| 34 | imputation | removal of outliers | removal of outliers | imputation | imputation | 86.96 | 52.24 | 98.73 |
| 35 | imputation | removal of outliers | imputation | imputation | imputation | 84.84 | 40 | 100 |
| 36 | imputation | imputation | removal of outliers | imputation | imputation | 86.11 | 45.99 | 99.51 |
| 37 | imputation | imputation | imputation | imputation | imputation | 85.81 | 44.44 | 99.54 |
| 38 | removal | log | removal of outliers | imputation | imputation | 84.19 | 40 | 98.77 |
| 39 | removal | log | imputation | imputation | imputation | 84.97 | 42.55 | 98.84 |
| 40 | removal | removal of outliers | removal of outliers | imputation | imputation | 86.36 | 53.38 | 97.47 |
| 41 | removal | removal of outliers | imputation | imputation | imputation | 86.96 | 47.83 | 100 |
| 42 | removal | imputation | removal of outliers | imputation | imputation | 84.19 | 40 | 98.77 |
| 43 | removal | imputation | imputation | imputation | imputation | 84.97 | 42.55 | 98.84 |

*Decision Tree:*

(With Gini Index):

| Dataset | Loan amount | Property value | Income | LTV | Rate of interest | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|---|---|---|---|
| 1 | - | - | - | - | - | 88.58 | 100 | 84.79 |
| 2 | - | log | removal of outliers | | | 89.58 | 100 | 86.1 |
| 3 | sqrt | log | imputation | | | 88.58 | 100 | 84.79 |
| 4 | - | removal of outliers | removal of outliers | | | 87.71 | 100 | 83.54 |
| 5 | sqrt | removal of outliers | imputation | | | 91.88 | 100 | 89.13 |
| 6 | - | imputation | removal of outliers | | | 89.58 | 100 | 86.1 |
| 7 | sqrt | imputation | imputation | | | 88.58 | 100 | 84.79 |
| 8 | log | log | imputation | | | 88.58 | 100 | 84.79 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 9 | log | removal of outliers | imputation | | | 91.88 | 100 | 89.13 |
| 10 | log | imputation | imputation | | | 88.58 | 100 | 84.79 |
| 11 | imputation | log | removal of outliers | | | 88.85 | 78.83 | 92.2 |
| 12 | imputation | log | imputation | | | 88.58 | 100 | 84.79 |
| 13 | imputation | removal of outliers | removal of outliers | | | 87.71 | 100 | 83.54 |
| 14 | imputation | removal of outliers | imputation | | | 91.88 | 100 | 89.13 |
| 15 | imputation | imputation | removal of outliers | | | 89.58 | 100 | 86.1 |
| 16 | imputation | imputation | imputation | | | 88.68 | 100 | 84.79 |
| 17 | removal | log | removal of outliers | | | 89.52 | 100 | 86.06 |
| 18 | removal | log | imputation | | | 87.59 | 100 | 83.53 |
| 19 | removal | removal of outliers | removal of outliers | | | 86.74 | 100 | 82.28 |
| 20 | removal | removal of outliers | imputation | | | 89.67 | 100 | 86.23 |
| 21 | removal | imputation | removal of outliers | | | 89.52 | 100 | 86.06 |
| 22 | removal | imputation | imputation | | | 87.59 | 100 | 83.53 |
| 23 | - | log | removal of outliers | imputation | imputation | 88.85 | 78.83 | 92.2 |
| 24 | sqrt | log | imputation | imputation | imputation | 86.33 | 53.47 | 97.24 |
| 25 | - | removal of outliers | removal of outliers | imputation | imputation | 88.09 | 76.87 | 91.9 |
| 26 | sqrt | removal of outliers | imputation | imputation | imputation | 88.63 | 63.57 | 97.1 |
| 27 | - | imputation | removal of outliers | imputation | imputation | 88.85 | 78.83 | 92.2 |
| 28 | sqrt | imputation | imputation | imputation | imputation | 86.33 | 53.47 | 97.24 |
| 29 | log | log | imputation | imputation | imputation | 86.33 | 53.47 | 97.24 |
| 30 | log | removal of outliers | imputation | imputation | imputation | 88.63 | 63.57 | 97.1 |
| 31 | log | imputation | imputation | imputation | imputation | 86.33 | 53.47 | 97.24 |
| 32 | imputation | log | removal of outliers | imputation | imputation | 88.85 | 78.83 | 92.2 |
| 33 | imputation | log | imputation | imputation | imputation | 86.33 | 53.47 | 97.24 |
| 34 | imputation | removal of outliers | removal of outliers | imputation | imputation | 88.09 | 76.87 | 91.9 |
| 35 | imputation | removal of outliers | imputation | imputation | imputation | 88.63 | 63.57 | 97.1 |
| 36 | imputation | imputation | removal of outliers | imputation | imputation | 88.85 | 78.83 | 92.2 |
| 37 | imputation | imputation | imputation | imputation | imputation | 86.33 | 53.47 | 97.24 |
| 38 | removal | log | removal of outliers | imputation | imputation | 86.58 | 65.93 | 93.4 |
| 39 | removal | log | imputation | imputation | imputation | 86.89 | 68.09 | 93.04 |
| 40 | removal | removal of outliers | removal of outliers | imputation | imputation | 88.83 | 75.19 | 93.42 |
| 41 | removal | removal of outliers | imputation | imputation | imputation | 89.13 | 79.71 | 92.27 |
| 42 | removal | imputation | removal of outliers | imputation | imputation | 86.58 | 65.93 | 93.4 |
| 43 | removal | imputation | imputation | imputation | imputation | 86.89 | 68.09 | 93.04 |

(With Information Gain):

| Dataset | Loan amount | Property value | Income | LTV | Rate of interest | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|---|---|---|---|
| 1 | - | - | - | - | - | 88.58 | 100 | 84.79 |
| 2 | - | log | removal of outliers | | | 89.58 | 100 | 86.1 |
| 3 | sqrt | log | imputation | | | 88.58 | 100 | 84.79 |
| 4 | - | removal of outliers | removal of outliers | | | 87.71 | 100 | 83.54 |
| 5 | sqrt | removal of outliers | imputation | | | 92.78 | 88.57 | 94.2 |
| 6 | - | imputation | removal of outliers | | | 89.58 | 100 | 86.1 |
| 7 | sqrt | imputation | imputation | | | 88.58 | 100 | 84.79 |
| 8 | log | log | imputation | | | 88.58 | 100 | 84.79 |
| 9 | log | removal of outliers | imputation | | | 92.78 | 88.57 | 94.2 |
| 10 | log | imputation | imputation | | | 88.58 | 100 | 84.79 |
| 11 | imputation | log | removal of outliers | | | 88.85 | 81.02 | 91.46 |
| 12 | imputation | log | imputation | | | 88.58 | 100 | 84.79 |
| 13 | imputation | removal of outliers | removal of outliers | | | 87.71 | 100 | 83.54 |
| 14 | imputation | removal of outliers | imputation | | | 92.78 | 88.57 | 94.2 |
| 15 | imputation | imputation | removal of outliers | | | 89.58 | 100 | 86.1 |
| 16 | imputation | imputation | imputation | | | 88.58 | 100 | 84.79 |
| 17 | removal | log | removal of outliers | | | 89.52 | 100 | 86.06 |
| 18 | removal | log | imputation | | | 86.89 | 83.69 | 87.94 |
| 19 | removal | removal of outliers | removal of outliers | | | 90.15 | 92.48 | 89.37 |
| 20 | removal | removal of outliers | imputation | | | 89.67 | 100 | 86.23 |
| 21 | removal | imputation | removal of outliers | | | 89.52 | 100 | 86.06 |
| 22 | removal | imputation | imputation | | | 86.89 | 83.69 | 87.94 |
| 23 | - | log | removal of outliers | imputation | imputation | 88.85 | 81.02 | 91.46 |
| 24 | sqrt | log | imputation | imputation | imputation | 86.33 | 53.47 | 97.24 |
| 25 | - | removal of outliers | removal of outliers | imputation | imputation | 88.09 | 76.87 | 91.9 |
| 26 | sqrt | removal of outliers | imputation | imputation | imputation | 88.63 | 63.57 | 97.1 |
| 27 | - | imputation | removal of outliers | imputation | imputation | 88.85 | 81.02 | 91.46 |
| 28 | sqrt | imputation | imputation | imputation | imputation | 86.33 | 53.47 | 97.24 |
| 29 | log | log | imputation | imputation | imputation | 86.33 | 53.47 | 97.24 |
| 30 | log | removal of outliers | imputation | imputation | imputation | 88.63 | 63.57 | 97.1 |
| 31 | log | imputation | imputation | imputation | imputation | 86.33 | 53.47 | 97.24 |
| 32 | imputation | log | removal of outliers | imputation | imputation | 89.03 | 81.02 | 91.71 |
| 33 | imputation | log | imputation | imputation | imputation | 86.33 | 53.47 | 97.24 |
| 34 | imputation | removal of outliers | removal of outliers | imputation | imputation | 88.09 | 76.87 | 91.9 |
| 35 | imputation | removal of outliers | imputation | imputation | imputation | 88.63 | 63.57 | 97.1 |
| 36 | imputation | imputation | removal of outliers | imputation | imputation | 89.03 | 81.02 | 91.71 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 37 | imputation | imputation | imputation | imputation | imputation | 86.33 | 53.47 | 97.24 |
| 38 | removal | log | removal of outliers | imputation | imputation | 86.58 | 65.93 | 93.4 |
| 39 | removal | log | imputation | imputation | imputation | 87.06 | 61.7 | 95.36 |
| 40 | removal | removal of outliers | removal of outliers | imputation | imputation | 88.83 | 75.19 | 93.42 |
| 41 | removal | removal of outliers | imputation | imputation | imputation | 89.13 | 78.99 | 92.51 |
| 42 | removal | imputation | removal of outliers | imputation | imputation | 86.58 | 65.93 | 93.4 |
| 43 | removal | imputation | imputation | imputation | imputation | 87.06 | 61.7 | 95.36 |

*Random forest:*

| Dataset | Loan amount | Property value | Income | LTV | Rate of interest | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|---|---|---|---|
| 1 | - | - | - | - | - | 87.72 | 81.94 | 89.63 |
| 2 | - | log | removal of outliers | | | 92.5 | 87.59 | 94.15 |
| 3 | sqrt | log | imputation | | | 88.06 | 82.64 | 89.86 |
| 4 | - | removal of outliers | removal of outliers | | | 90.17 | 86.57 | 91.39 |
| 5 | sqrt | removal of outliers | imputation | | | 92.06 | 87.86 | 93.48 |
| 6 | - | imputation | removal of outliers | | | 92.87 | 87.59 | 94.63 |
| 7 | sqrt | imputation | imputation | | | 87.89 | 82.64 | 89.63 |
| 8 | log | log | imputation | | | 88.06 | 82.64 | 89.86 |
| 9 | log | removal of outliers | imputation | | | 92.42 | 86.43 | 94.44 |
| 10 | log | imputation | imputation | | | 88.58 | 83.33 | 90.32 |
| 11 | imputation | log | removal of outliers | | | 89.95 | 78.1 | 93.9 |
| 12 | imputation | log | imputation | | | 87.72 | 81.94 | 89.63 |
| 13 | imputation | removal of outliers | removal of outliers | | | 89.41 | 85.07 | 90.89 |
| 14 | imputation | removal of outliers | imputation | | | 92.24 | 87.86 | 93.72 |
| 15 | imputation | imputation | removal of outliers | | | 92.5 | 87.59 | 94.15 |
| 16 | imputation | imputation | imputation | | | 88.06 | 83.33 | 89.63 |
| 17 | removal | log | removal of outliers | | | 89.89 | 81.48 | 92.67 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 18 | removal | log | imputation | | | 88.81 | 80.85 | 91.42 |
| 19 | removal | removal of outliers | removal of outliers | | | 90.91 | 88.72 | 91.65 |
| 20 | removal | removal of outliers | imputation | | | 91.67 | 87.68 | 93 |
| 21 | removal | imputation | removal of outliers | | | 90.81 | 81.48 | 93.89 |
| 22 | removal | imputation | imputation | | | 89.51 | 81.56 | 92.11 |
| 23 | - | log | removal of outliers | imputation | imputation | 89.95 | 81.75 | 92.68 |
| 24 | sqrt | log | imputation | imputation | imputation | 88.06 | 75.69 | 92.17 |
| 25 | - | removal of outliers | removal of outliers | imputation | imputation | 88.47 | 82.09 | 90.63 |
| 26 | sqrt | removal of outliers | imputation | imputation | imputation | 91.52 | 82.14 | 94.69 |
| 27 | - | imputation | removal of outliers | imputation | imputation | 90.49 | 79.56 | 94.15 |
| 28 | sqrt | imputation | imputation | imputation | imputation | 87.37 | 72.92 | 92.17 |
| 29 | log | log | imputation | imputation | imputation | 87.2 | 74.31 | 91.47 |
| 30 | log | removal of outliers | imputation | imputation | imputation | 90.16 | 78.57 | 94.69 |
| 31 | log | imputation | imputation | imputation | imputation | 87.02 | 75 | 91.01 |
| 32 | imputation | log | removal of outliers | imputation | imputation | 90.13 | 79.56 | 93.66 |
| 33 | imputation | log | imputation | imputation | imputation | 87.72 | 77.08 | 91.24 |
| 34 | imputation | removal of outliers | removal of outliers | imputation | imputation | 88.66 | 80.6 | 91.39 |
| 35 | imputation | removal of outliers | imputation | imputation | imputation | 90.43 | 77.14 | 94.93 |
| 36 | imputation | imputation | removal of outliers | imputation | imputation | 90.68 | 79.56 | 94.39 |
| 37 | imputation | imputation | imputation | imputation | imputation | 86.85 | 73.61 | 91.24 |
| 38 | removal | log | removal of outliers | imputation | imputation | 88.05 | 81.56 | 93.64 |
| 39 | removal | log | imputation | imputation | imputation | 88.11 | 74.47 | 92.58 |
| 40 | removal | removal of outliers | removal of outliers | imputation | imputation | 89.93 | 80.45 | 92.41 |
| 41 | removal | removal of outliers | imputation | imputation | imputation | 91.3 | 84.78 | 93.48 |
| 42 | removal | imputation | removal of outliers | imputation | imputation | 88.6 | 71.85 | 94.13 |
| 43 | removal | imputation | imputation | imputation | imputation | 88.11 | 75.89 | 92.11 |

## RESULTS

We want to predict whether a loan will be approved or not, here the cost of a false positive is low and we want to capture as many positive approvals as possible; even if we identify a few false approvals. And hence we will look for sensitivity instead of specificity in our models as to compare them.

So, we have analyzed all the created models and shortlisted the ones with maximum accuracy and sensitivity from all the methods:

*Logistic regression:*

| Dataset | Loan amount | Property value | Income | LTV | Rate of interest | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|---|---|---|---|

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 4 | - | removal of outliers | removal of outliers | | | 87.33 | 51.49 | 99.49 |

```
> confusionMatrix(pL4_glm, testing4$Status,positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 393  65
         1   2  69

               Accuracy : 0.8733
                 95% CI : (0.842, 0.9005)
    No Information Rate : 0.7467
    P-Value [Acc > NIR] : 4.860e-13

                  Kappa : 0.6036

 Mcnemar's Test P-Value : 3.605e-14

            Sensitivity : 0.5149
            Specificity : 0.9949
         Pos Pred Value : 0.9718
         Neg Pred Value : 0.8581
             Prevalence : 0.2533
         Detection Rate : 0.1304
   Detection Prevalence : 0.1342
      Balanced Accuracy : 0.7549

       'Positive' Class : 1
```

Model made using all the independent variables of dataset '4' is best among all the 84 models of logistic regression with an accuracy of 87.33% and sensitivity of 51.49%.

*Naïve Bayes:*

| Dataset | Loan amount | Property value | Income | LTV | Rate of interest | LA Accuracy | LA Sensitivity | LA Specificity |
|---|---|---|---|---|---|---|---|---|
| 20 | removal | removal of outliers | imputation | | | 87.86 | 52.9 | 99.52 |
| 19 | removal | removal of outliers | removal of outliers | | | 86.93 | 54.14 | 97.97 |

```
> confusionMatrix(pL20_NBLA, testing20$Status,positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 412  65
         1   2  73

               Accuracy : 0.8786
                 95% CI : (0.8484, 0.9047)
    No Information Rate : 0.75
    P-Value [Acc > NIR] : 4.256e-14

                  Kappa : 0.6182

 Mcnemar's Test P-Value : 3.605e-14

            Sensitivity : 0.5290
            Specificity : 0.9952
         Pos Pred Value : 0.9733
         Neg Pred Value : 0.8637
             Prevalence : 0.2500
         Detection Rate : 0.1322
   Detection Prevalence : 0.1359
      Balanced Accuracy : 0.7621

       'Positive' Class : 1

> confusionMatrix(pL19_NBLA, testing19$Status,positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 387  61
         1   8  72

               Accuracy : 0.8693
                 95% CI : (0.8375, 0.8969)
    No Information Rate : 0.7481
    P-Value [Acc > NIR] : 4.946e-12

                  Kappa : 0.6005

 Mcnemar's Test P-Value : 3.848e-10

            Sensitivity : 0.5414
            Specificity : 0.9797
         Pos Pred Value : 0.9000
         Neg Pred Value : 0.8638
             Prevalence : 0.2519
         Detection Rate : 0.1364
   Detection Prevalence : 0.1515
      Balanced Accuracy : 0.7606

       'Positive' Class : 1
```

| Dataset | Loan amount | Property value | Income | LTV | Rate of interest | PV Accuracy | PV Sensitivity | PV Specificity |
|---|---|---|---|---|---|---|---|---|
| 20 | removal | removal of outliers | imputation | | | 87.86 | 52.9 | 99.52 |
| 19 | removal | removal of outliers | removal of outliers | | | 87.12 | 54.14 | 98.23 |

```
> confusionMatrix(pL20_NBPV, testing20$Status,positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 412  65
         1   2  73

               Accuracy : 0.8786
                 95% CI : (0.8484, 0.9047)
    No Information Rate : 0.75
    P-Value [Acc > NIR] : 4.256e-14

                  Kappa : 0.6182

 Mcnemar's Test P-Value : 3.605e-14

            Sensitivity : 0.5290
            Specificity : 0.9952
         Pos Pred Value : 0.9733
         Neg Pred Value : 0.8637
             Prevalence : 0.2500
         Detection Rate : 0.1322
   Detection Prevalence : 0.1359
      Balanced Accuracy : 0.7621

       'Positive' Class : 1
```

```
> confusionMatrix(pL19_NBPV, testing19$Status,positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 388  61
         1   7  72

               Accuracy : 0.8712
                 95% CI : (0.8396, 0.8986)
    No Information Rate : 0.7481
    P-Value [Acc > NIR] : 2.175e-12

                  Kappa : 0.6051

 Mcnemar's Test P-Value : 1.300e-10

            Sensitivity : 0.5414
            Specificity : 0.9823
         Pos Pred Value : 0.9114
         Neg Pred Value : 0.8641
             Prevalence : 0.2519
         Detection Rate : 0.1364
   Detection Prevalence : 0.1496
      Balanced Accuracy : 0.7618

       'Positive' Class : 1
```

When it comes to accuracy, both the models (Loan amount/Property value) made using dataset '20' are best with an accuracy of 87.86%.

When it comes to sensitivity, both the models (Loan amount/Property value) made using dataset '19' are best with a sensitivity of 54.14%.

But when it comes to both model (Property value) made using dataset '19' is best with an accuracy of 87.12% and a sensitivity of 54.14%.

*Decision tree:*

Gini Index:

| Dataset | Loan amount | Property value | Income | LTV | Rate of interest | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|---|---|---|---|
| 5 | sqrt | removal of outliers | imputation | | | 91.88 | 100 | 89.13 |
| 9 | log | removal of outliers | imputation | | | 91.88 | 100 | 89.13 |
| 14 | imputation | removal of outliers | imputation | | | 91.88 | 100 | 89.13 |

```
> confusionMatrix(pL5_gini, testing5$Status,positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 369   0
         1  45 140

               Accuracy : 0.9188
                 95% CI : (0.8928, 0.9401)
    No Information Rate : 0.7473
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8056

 Mcnemar's Test P-Value : 5.412e-11

            Sensitivity : 1.0000
            Specificity : 0.8913
         Pos Pred Value : 0.7568
         Neg Pred Value : 1.0000
             Prevalence : 0.2527
         Detection Rate : 0.2527
   Detection Prevalence : 0.3339
      Balanced Accuracy : 0.9457

       'Positive' Class : 1
```

Information Gain:

| Dataset | Loan amount | Property value | Income | LTV | Rate of interest | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|---|---|---|---|
| 5 | sqrt | removal of outliers | imputation | | | 92.78 | 88.57 | 94.2 |
| 9 | log | removal of outliers | imputation | | | 92.78 | 88.57 | 94.2 |
| 14 | imputation | removal of outliers | imputation | | | 92.78 | 88.57 | 94.2 |

```
> confusionMatrix(pL5_info, testing5$Status,positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 390  16
         1  24 124

               Accuracy : 0.9278
                 95% CI : (0.903, 0.9479)
    No Information Rate : 0.7473
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8124

 Mcnemar's Test P-Value : 0.2684

            Sensitivity : 0.8857
            Specificity : 0.9420
         Pos Pred Value : 0.8378
         Neg Pred Value : 0.9606
             Prevalence : 0.2527
         Detection Rate : 0.2238
   Detection Prevalence : 0.2671
      Balanced Accuracy : 0.9139

       'Positive' Class : 1
```

Models made using information gain method with datasets ‘5’, ‘9’ and ‘14’ are the best when it comes to accuracy with an accuracy of 92.78%.

But when it comes to both accuracy and sensitivity, models made using gini index method with datasets ‘5’, ‘9’ and ‘14’ are best with an accuracy of 91.88% and a sensitivity of 100%.

*Random forest:*

| Dataset | Loan amount | Property value | Income | LTV | Rate of interest | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|---|---|---|---|
| 6 | - | imputation | removal of outliers | | | 92.87 | 87.59 | 94.63 |
| 19 | removal | removal of outliers | removal of outliers | | | 90.91 | 88.72 | 91.65 |

```
confusionMatrix(pL6_rf, testing6$Status,positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 388  17
         1  22 120

               Accuracy : 0.9287
                 95% CI : (0.9038, 0.9488)
    No Information Rate : 0.7495
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8124

 Mcnemar's Test P-Value : 0.5218

            Sensitivity : 0.8759
            Specificity : 0.9463
         Pos Pred Value : 0.8451
         Neg Pred Value : 0.9580
             Prevalence : 0.2505
         Detection Rate : 0.2194
   Detection Prevalence : 0.2596
      Balanced Accuracy : 0.9111

       'Positive' Class : 1


confusionMatrix(p19_rf, testing19$Status,positive = "1")
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 362  15
         1  33 118

               Accuracy : 0.9091
                 95% CI : (0.8813, 0.9322)
    No Information Rate : 0.7481
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.7692

 Mcnemar's Test P-Value : 0.01414

            Sensitivity : 0.8872
            Specificity : 0.9165
         Pos Pred Value : 0.7815
         Neg Pred Value : 0.9602
             Prevalence : 0.2519
         Detection Rate : 0.2235
   Detection Prevalence : 0.2860
      Balanced Accuracy : 0.9018

       'Positive' Class : 1
```

When it comes to accuracy, model made using dataset ‘6’ is best with an accuracy of 92.87%.

But when it comes to sensitivity, model made using dataset ‘19’ is best with a sensitivity of 88.72%.

**CONCLUSION**

When it comes to accuracy, the best model is obtained from dataset ‘6’ using random forest method with highest accuracy of 92.87%.

But when it comes to the best model for our objective, it is obtained from dataset ‘5’, ‘9’ and ‘14’ using decision tree method with gini index split, having an accuracy of 91.88% and a sensitivity of 100%.

But as best model is from random forest, we will use its output to evaluate the importance of independent variables on our dependent variable.

```
> varImp(modelL6_rf)
rf variable importance

  only 20 most important variables shown (out of 30)

                        Overall
rate_of_interest        100.000
credit_typeEQUI          75.541
LTV                      21.176
income                   18.308
property_value           16.343
Credit_Score             12.781
loan_amount              11.384
age35-44                  2.198
loan_purposep3            1.819
credit_typeCRIF           1.675
GenderMale                1.618
approv_in_advpre          1.595
loan_purposep4            1.583
credit_typeEXP            1.575
age45-54                  1.554
GenderSex Not Available   1.524
GenderJoint               1.522
Regionsouth               1.300
RegionNorth               1.259
age55-64                  1.219
```

Hence, rate_of_interest is most important predictor of status followed by other predictors in descending order.

**APPENDIX**

set.seed() value was  used as hundred for all the models training to testing data for

each ML model was taken as 80:20

https://datascience.stackexchange.com/questions/9087/correlation-and-naive-bayes

https://stats.stackexchange.com/questions/409094/can-a-prediction-be-better-with-insignificantvariables-than-with-only-significa

https://datascience.stackexchange.com/questions/113403/how-much-percentage-of-outliers-are-allowedin-a-data