MACHINE LEARNING PROJECT

IDENTIFICATION OF THE 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.
- After going through these rigorous processes, 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

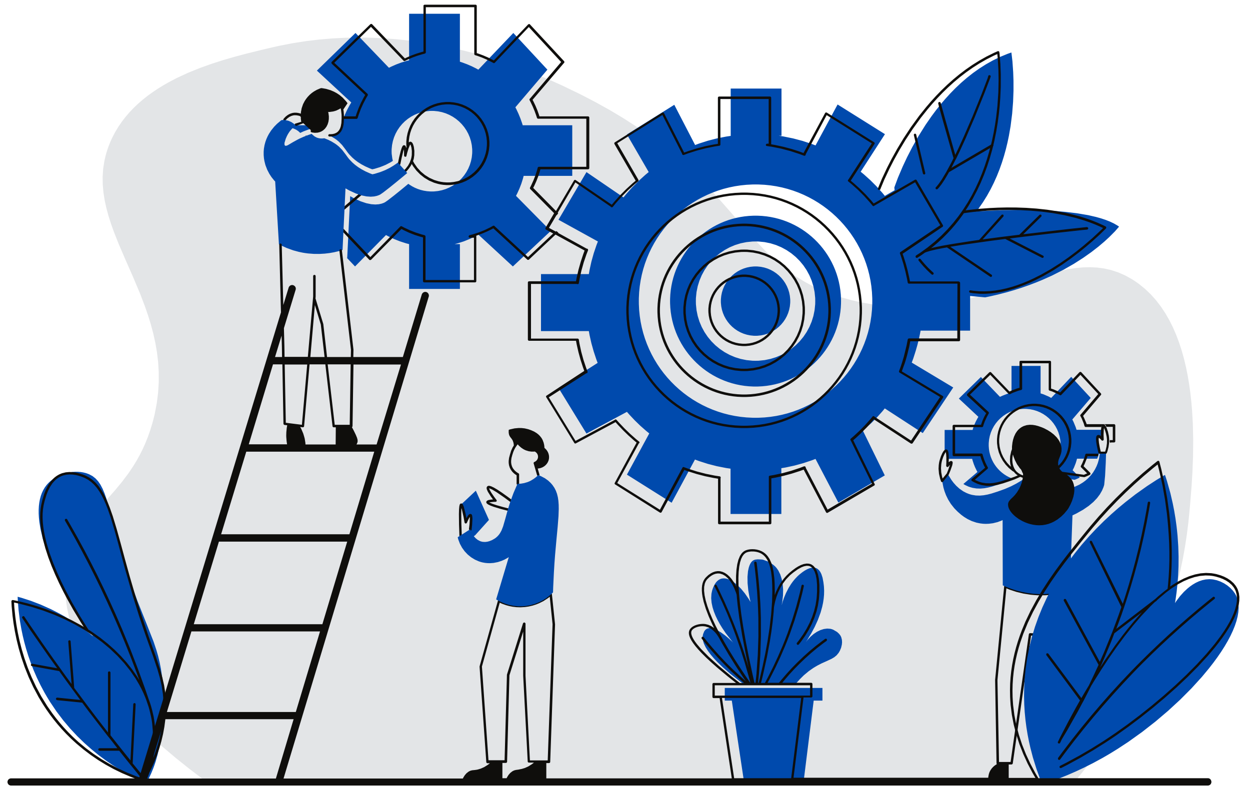
The dataset is named as '**LoanDefaulters**' and contains past records of people who applied for loan with all their information and whether their loan was approved or not approved.

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 (l1 or l2)
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)

Discrete, Categorical, Continuous

OBJECTIVES

- 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.
- 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.
- We will only use imputation of missing values and not removal of them to avoid making the dataset smaller.

IDENTIFICATION OF VARIABLES

- 'Status' will be the **dependent variable**.
- 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 **rest** are **categorical** variables.

IDENTIFIED ISSUES AND POSSIBLE TREATMENT

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 two have moderate outliers: rate_of_interest (109 outliers) and LTV (111 outliers)

- 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 , so we will make two categories of models;
  1. outliers are not treated
  2. outlier values are treated using imputation.

	square root transformation	Log transformation	Removal	Imputation
Loan Amount (32 outliers present)	17 (outliers still present) Normality achieved	17 (outliers still present) Normality achieved	0 (outliers present) Normality achieved	0 (outliers present) Normality achieved
Property value (121 outliers present)	Not possible	58 (outliers still present)	25 (outliers still present) Normality achieved	54 (outliers still present) Normality achieved
Income (152 outliers present)	Not possible	Not possible	47 (outliers still present) Normality achieved	72 (outliers still present) Normality achieved

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		

POSSIBLE DATASETS WITH DIFFERENT TREATMENT

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

PREDICTIVE MODELS

POSSIBLE METHODS

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

POSSIBLE MODELS

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**

RESULTS

So, we have analyzed all the created models and shortlisted the ones with maximum accuracy and sensitivity(the cost of a false positive is low and we want to capture as many positive approvals as possible) from all the methods

LOGISTIC REGRESSION

Model No.	Loan amount	Property value	Income	LTV	Rate of interest
4	-	removal of outliers	removal of outliers		

Accuracy	Sensitivity	Specificity
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
```

NAÏVE BAYES

Dataset	Loan amount	Property value	Income	LTV	Rate of interest
20	removal	removal of outliers	imputation		
19	removal	removal of outliers	removal of outliers		

LA Accuracy	LA Sensitivity	LA Specificity
87.86	52.9	99.52
86.93	54.14	97.97

```
> 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
```

NAÏVE BAYES

Dataset	Loan amount	Property value	Income	LTV	Rate of interest
20	removal	removal of outliers	imputation		
19	removal	removal of outliers	removal of outliers		

LA Accuracy	LA Sensitivity	LA Specificity
87.86	52.9	99.52
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
```

NAÏVE BAYES

Dataset	Loan amount	Property value	Income	LTV	Rate of interest
20	removal	removal of outliers	imputation		
19	removal	removal of outliers	removal of outliers		

PV Accuracy	PV Sensitivity	PV Specificity
87.86	52.9	99.52
87.12	54.14	98.23

```
> 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
```

NAÏVE BAYES

Dataset	Loan amount	Property value	Income	LTV	Rate of interest
20	removal	removal of outliers	imputation		
19	removal	removal of outliers	removal of outliers		

PV Accuracy	PV Sensitivity	PV Specificity
87.86	52.9	99.52
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
```

DECISION TREE

Gini Index:

Dataset	Loan amount	Property value	Income	LTV	Rate of interest
5	sqrt	removal of outliers	imputation		
9	log	removal of outliers	imputation		
14	imputation	removal of outliers	imputation		

Accuracy	Sensitivity	Specificity
91.88	100	89.13
91.88	100	89.13
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
```

DECISION TREE

Information Gain:

Dataset	Loan amount	Property value	Income	LTV	Rate of interest
5	sqrt	removal of outliers	imputation		
9	log	removal of outliers	imputation		
14	imputation	removal of outliers	imputation		

Accuracy	Sensitivity	Specificity
92.78	88.57	94.2
92.78	88.57	94.2
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
```

RANDOM FOREST

Dataset	Loan amount	Property value	Income	LTV	Rate of interest
6	-	imputation	removal of outliers		
19	removal	removal of outliers	removal of outliers		

Accuracy	Sensitivity	Specificity
92.87	87.59	94.63
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
```

RANDOM FOREST

Dataset	Loan amount	Property value	Income	LTV	Rate of interest
6	-	imputation	removal of outliers		
19	removal	removal of outliers	removal of outliers		

Accuracy	Sensitivity	Specificity
92.87	87.59	94.63
90.91	88.72	91.65

```
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
```

CONCLUSION

- In terms of accuracy, the best model is obtained from dataset '6' using random forest method with highest accuracy of 92.87%.
- In terms of sensitivity that is our objective, the best model 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 the 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
```

THANK YOU