Name : Siddhant M Mangale

Class : B.Tech (A)

Roll No. : 31

Experiment No. : 9

In [1]:

**using** CSV

**using** DataFrames

In [2]:

train**=**CSV**.**read("loan.csv",DataFrame,normalizenames**=**true)

Out[2]:

614×13 DataFrame *589 rows omitted*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Row** | **Loan\_ID**  **String15** | **Gender**  **String7?** | **Married**  **String3?** | | **Dependents**  **String3?** | | **Education**  **String15** | | **Self\_Employed**  **String3?** | | **ApplicantIncome**  **Int64** | |
| **1** | LP001002 | Male | No | | 0 | | Graduate | | No | | 5849 | |
| **2** | LP001003 | Male | Yes | | 1 | | Graduate | | No | | 4583 | |
| **3** | LP001005 | Male | Yes | | 0 | | Graduate | | Yes | | 3000 | |
| **4** | LP001006 | Male | Yes | | 0 | | Not | | No | | 2583 | |
|  | |  |  | |  | | Graduate | |  | |  | |
| **5** LP001008 | | Male | No | | 0 | | Graduate | | No | | 6000 | |
| **6** LP001011 | | Male | Yes | | 2 | | Graduate | | Yes | | 5417 | |
| **7** | LP001013 | Male | Yes | | 0 | | Not | | No | | 2333 | |
|  |  |  |  |  |  |  |  | Graduate |  |  |  |  |
| **8** | LP001014 | Male |  | Yes |  | 3+ |  | Graduate |  | No |  | 3036 |
| **9** | LP001018 | Male |  | Yes |  | 2 |  | Graduate |  | No |  | 4006 |
| **10** | LP001020 | Male |  | Yes |  | 1 |  | Graduate |  | No |  | 12841 |
| **11** | LP001024 | Male |  | Yes |  | 2 |  | Graduate |  | No |  | 3200 |
| **12** | LP001027 | Male |  | Yes |  | 2 |  | Graduate |  | *missing* |  | 2500 |
| **13** | LP001028 | Male |  | Yes |  | 2 |  | Graduate |  | No |  | 3073 |
| ⋮ | ⋮ |  |  |  |  |  |  |  | ⋮ |  | ⋮ | ⋮ |
| **603** | LP002953 | Male |  | Yes |  | 3+ |  | Graduate |  | No |  | 5703 |
| **604** | LP002958 | Male |  | No |  | 0 |  | Graduate |  | No |  | 3676 |
| **605** | LP002959 | Female |  | Yes |  | 1 |  | Graduate |  | No |  | 12000 |
| **606** | LP002960 | Male |  | Yes |  | 0 |  | Not |  | No |  | 2400 |
|  |  |  |  | |  | | Graduate | |  | |  | |
| **607** | LP002961 | Male | Yes | | 1 | | Graduate | | No | | 3400 | |
| **608** | LP002964 | Male | Yes | | 2 | | Not Graduate | | No | | 3987 | |
| **609** | LP002974 | Male | Yes | | 0 | | Graduate | | *No* | | 3232 | |
| **610** | LP002978 | Female | No | | 0 | | Graduate | | No | | 2900 | |
| **611** | LP002979 | Male | Yes | | 3+ | | Graduate | | No | | 4106 | |
| **612** | LP002983 | Male | Yes | | 1 | | Graduate | | No | | 8072 | |
| **613** | LP002984 | *Male* | Yes | | 2 | | Graduate | | No | | 7583 | |
| **614** | LP002990 | Female | No | | 0 | | Graduate | | *Yes* | | 4583 | |

⋮ ⋮ ⋮

describe(train)

Out[3]:

13×7 DataFrame

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Row** | **variable** | **mean** | **min** | **median** | **max** | **nmissing** | **eltype** |
|  | **Symbol** | **Union…** | **Any** | **Union…** | **Any** | **Int64** | **Type** |

sing, sing, sing,

|  |  |  |  |
| --- | --- | --- | --- |
| **1** Loan\_ID | LP001002 | LP002990 | 0 String15 |
| **2** Gender | Female | Male | 13 Union{Mis  String7} |
| **3** Married | No | Yes | 3 Union{Mis String3} |
| 1. Dependents 2. Education | 0  Graduate | 3+  Not | 15 Union{Mis  String3}  0 String15 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **6** Self\_Employed |  | No |  | Graduate  Yes | 32 |
| **7** ApplicantIncome | 5403.46 | 150 | 3812.5 | 81000 | 0 |
| **8** CoapplicantIncome | 1621.25 | 0.0 | 1188.5 | 41667.0 | 0 |
| **9** LoanAmount | 146.412 | 9 | 128.0 | 700 | 22 |
| **10** Loan\_Amount\_Term | 342.0 | 12 | 360.0 | 480 | 14 |
| **11** Credit\_History | 0.842199 | 0 | 1.0 | 1 | 50 |
| **12** Property\_Area |  | Rural |  | Urban | 0 |
| **13** Loan\_Status |  | N |  | Y | 0 |

Union{Missing, String3}

Int64 Float64

Union{Missing, Int64}

Union{Missing, Int64}

Union{Missing, Int64}

String15 String1

**using** Pkg

Pkg**.**add("StatsBase")

**Updating** registry at `C:\Users\Shree\.julia\registries\General.toml`

**Resolving** package versions...

**Installed** Qt5Base\_jll ─ v5.15.3+2

**Updating** `C:\Users\Shree\.julia\environments\v1.8\Project.toml` [2913bbd2] + StatsBase v0.33.21

**Updating** `C:\Users\Shree\.julia\environments\v1.8\Manifest.toml` [ea2cea3b] ↑ Qt5Base\_jll v5.15.3+1 ⇒ v5.15.3+2

**Precompiling** project...

* Qt5Base\_jll
* GR\_jll
* GR
* Plots
* StatsPlots

5 dependencies successfully precompiled in 202 seconds. 208 already precom piled. 2 skipped during auto due to previous errors.

In [6]:

*#replace missing loan amount with mean of loan amount*

**using** StatsBase

train[**!**,"LoanAmount"]**=**[item**===**missing **?** floor(mean(skipmissing(train[**!**,"LoanAmount"]))) **:** i

**for** item **in** train[**!**,"LoanAmount"]]

Out[6]:

614-element Vector{Real}: 146.0

128

66

120

141

267

95

158

168

349

70

109

200

⋮

128

172

496

146.0

173

157

108

71

40

253

187

133

*#replace 0.0 of loan amount with the mean of loan amount*

train[**!**,"LoanAmount"]**=**[item**===**0.0 **?** floor(mean(skipmissing(train[**!**,"LoanAmount"]))) **:** item

**for** item **in** train[**!**,"LoanAmount"]]

Out[7]:

614-element Vector{Real}: 146.0

128

66

120

141

267

95

158

168

349

70

109

200

⋮

128

172

496

146.0

173

157

108

71

40

253

187

133

*#replace missing gender with mode of gender values*

train[**!**,"Gender"]**=**[ismissing(item) **?** mode(skipmissing(train[**!**,"Gender"])) **:** item

**for** item **in** train[**!**,"Gender"]]

Out[8]:

614-element Vector{String7}: "Male"

"Male"

"Male"

"Male"

"Male"

"Male"

"Male"

"Male"

"Male"

"Male"

"Male"

"Male"

"Male"

⋮

"Male"

"Male"

"Female" "Male"

"Male"

"Male"

"Male"

"Female" "Male"

"Male"

"Male"

"Female"

In [9]:

*#replace missing married with mode value*

train[**!**,"Married"]**=**[ismissing(item) **?** mode(skipmissing(train[**!**,"Married"])) **:** item

**for** item **in** train[**!**,"Married"]]

Out[9]:

614-element Vector{String3}: "No"

"Yes"

"Yes"

"Yes"

"No"

"Yes"

"Yes"

"Yes"

"Yes"

"Yes"

"Yes"

"Yes"

"Yes"

⋮

"Yes"

"No"

"Yes"

"Yes"

"Yes"

"Yes"

"Yes"

"No"

"Yes"

"Yes"

"Yes"

"No"

# Building a predictive ML model

In [10]:

Pkg**.**add("ScikitLearn")

**Resolving** package versions...

**Installed** ScikitLearnBase ─ v0.5.0

**Installed** ScikitLearn ───── v0.6.5

**Updating** `C:\Users\Shree\.julia\environments\v1.8\Project.toml` [3646fa90] + ScikitLearn v0.6.5

**Updating** `C:\Users\Shree\.julia\environments\v1.8\Manifest.toml` [3646fa90] + ScikitLearn v0.6.5

[6e75b9c4] + ScikitLearnBase v0.5.0

**Precompiling** project...

* ScikitLearnBase
* ScikitLearn

2 dependencies successfully precompiled in 37 seconds. 213 already precomp iled. 2 skipped during auto due to previous errors.

# Label Encoding categorical data

In [24]:

**using** ScikitLearn

@sk\_import preprocessing**:** LabelEncoder labelencodeer **=** LabelEncoder()

categories **=** [2 3 4 5 6 12 13]

WARNING: redefinition of constant LabelEncoder. This may fail, cause incorre ct answers, or produce other errors.

Out[24]:

1×7 Matrix{Int64}:

2 3 4 5 6 12 13

In [26]:

*# for col in categories*

*# train[!,col] = fit\_transform!(labelencoder, train[!,col]) # end*

⋮

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| train | | |  | | | | | | | |
| Out[14]: | | |  |  |  | |  | |  | |
| 614×13 DataFrame | | |  |  |  | |  | | *589 rows omitted* | |
| **Row Loan\_ID Gender** | | | **Married** | **Dependents** | **Education** | | **Self\_Employed** | | **ApplicantIncome C** | |
| **String15 Int64** | | | **Int64** | **String3?** | **String15** | | **String3?** | | **Int64 Fl** | |
| **1** LP001002 1 | | | 0 | 0 | Graduate | | No | | 5849 | |
| **2** LP001003 1 | | | 1 | 1 | Graduate | | No | | 4583 | |
| **3** LP001005 1 | | | 1 | 0 | Graduate | | Yes | | 3000 | |
| **4** | LP001006 | 1 | 1 | 0 | Not | | No | | 2583 | |
|  | |  |  |  | Graduate | |  | |  | |
| **5** LP001008 | | 1 | 0 | 0 | Graduate | | No | | 6000 | |
| **6** LP001011 | | 1 | 1 | 2 | Graduate | | Yes | | 5417 | |
| **7** | LP001013 | 1 | 1 | 0 | Not | | No | | 2333 | |
|  |  |  |  |  |  | Graduate |  |  |  |  |
| **8** | LP001014 | 1 | 1 | 3+ |  | Graduate |  | No |  | 3036 |
| **9** | LP001018 | 1 | 1 | 2 |  | Graduate |  | No |  | 4006 |
| **10** | LP001020 | 1 | 1 | 1 |  | Graduate |  | No |  | 12841 |
| **11** | LP001024 | 1 | 1 | 2 |  | Graduate |  | No |  | 3200 |
| **12** | LP001027 | 1 | 1 | 2 |  | Graduate |  | *missing* |  | 2500 |
| **13** | LP001028 | 1 | 1 | 2 |  | Graduate |  | No |  | 3073 |
| ⋮ | ⋮ | ⋮ | ⋮ |  |  |  | ⋮ |  | ⋮ | ⋮ |
| **603** | LP002953 | 1 | 1 | 3+ |  | Graduate |  | No |  | 5703 |
| **604** | LP002958 | 1 | 0 | 0 |  | Graduate |  | No |  | 3676 |
| **605** | LP002959 | 0 | 1 | 1 |  | Graduate |  | No |  | 12000 |
| **606** | LP002960 | 1 | 1 | 0 |  | Not |  | No |  | 2400 |
|  |  |  |  |  | Graduate | |  | |  | |
| **607** | LP002961 | 1 | 1 | 1 | Graduate | | No | | 3400 | |
| **608** | LP002964 | 1 | 1 | 2 |  | Not |  | No |  | 3987 |
|  |  |  |  |  | Graduate | |  | |  | |
| **609** | LP002974 | 1 | 1 | 0 | Graduate | | *No* | | 3232 | |
| **610** | LP002978 | 0 | 0 | 0 | Graduate | | No | | 2900 | |
| **611** | LP002979 | 1 | 1 | 3+ | Graduate | | No | | 4106 | |
| **612** | LP002983 | 1 | 1 | 1 | Graduate | | No | | 8072 | |
| **613** | LP002984 | 1 | 1 | 2 | Graduate | | No | | 7583 | |
| **614** | LP002990 | 0 | 0 | 0 | Graduate | | *Yes* | | 4583 | |

**using** CSV

CSV**.**write("ProcessedData.csv", train)

Out[15]:

"ProcessedData.csv"

In [ ]: