Loan Approval Prediction

1. Install and load required modules

```
# instaling modules

install.packages(c("tidyverse", "dplyr", "tidyr"))
```

```
## Installing packages into '/Users/yaredshewarade/Library/R/arm64/4.3/library'
## (as 'lib' is unspecified)
```

```
##
## The downloaded binary packages are in
##  /var/folders/gs/yc2t8xm15hjfjrsfxy7tsz200000gn/T//Rtmp459jeZ/downloaded_packages
```

```
install.packages("caret")
```

```
## Installing package into '/Users/yaredshewarade/Library/R/arm64/4.3/library'
## (as 'lib' is unspecified)
```

```
##
## The downloaded binary packages are in
##  /var/folders/gs/yc2t8xm15hjfjrsfxy7tsz200000gn/T//Rtmp459jeZ/downloaded_packages
```

```
install.packages("tm")
```

```
## Installing package into '/Users/yaredshewarade/Library/R/arm64/4.3/library'
## (as 'lib' is unspecified)
```

```
##
## The downloaded binary packages are in
##  /var/folders/gs/yc2t8xm15hjfjrsfxy7tsz200000gn/T//Rtmp459jeZ/downloaded_packages
```

```
# Loading libraries
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(tm)
```

```
## Loading required package: NLP
##
## Attaching package: 'NLP'
##
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```r
library(dplyr)
library(tidyr)
```

2. Dataset loading

```r
# loading csv loan datset
loan_data <- read.csv("dataset.csv")
head(loan_data)
```

```
##   loan_id no_of_dependents     education self_employed income_annum loan_amount
## 1       1                2      Graduate            No      9600000    29900000
## 2       2                0  Not Graduate           Yes      4100000    12200000
## 3       3                3      Graduate            No      9100000    29700000
## 4       4                3      Graduate            No      8200000    30700000
## 5       5                5  Not Graduate           Yes      9800000    24200000
## 6       6                0      Graduate           Yes      4800000    13500000
##   loan_term cibil_score residential_assets_value commercial_assets_value
## 1        12         778                  2400000                17600000
## 2         8         417                  2700000                 2200000
## 3        20         506                  7100000                 4500000
## 4         8         467                 18200000                 3300000
## 5        20         382                 12400000                 8200000
## 6        10         319                  6800000                 8300000
##   luxury_assets_value bank_asset_value loan_status
## 1            22700000          8000000    Approved
## 2             8800000          3300000    Rejected
## 3            33300000         12800000    Rejected
## 4            23300000          7900000    Rejected
## 5            29400000          5000000    Rejected
## 6            13700000          5100000    Rejected
```

3. Data cleaning

```r
# view the structure of the data
str(loan_data)
```

```
## 'data.frame':    4269 obs. of  13 variables:
##  $ loan_id                 : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ no_of_dependents        : int  2 0 3 3 5 0 5 2 0 5 ...
##  $ education               : chr  " Graduate" " Not Graduate" " Graduate" " Graduate" ...
##  $ self_employed           : chr  " No" " Yes" " No" " No" ...
##  $ income_annum            : int  9600000 4100000 9100000 8200000 9800000 4800000 8700000 5700000 80
##  $ loan_amount             : int  29900000 12200000 29700000 30700000 24200000 13500000 33000000 150
```

```
##  $ loan_term               : int  12 8 20 8 20 10 4 20 20 10 ...
##  $ cibil_score             : int  778 417 506 467 382 319 678 382 782 388 ...
##  $ residential_assets_value: int  2400000 2700000 7100000 18200000 12400000 6800000 22500000 1320000
##  $ commercial_assets_value : int  17600000 2200000 4500000 3300000 8200000 8300000 14800000 5700000 
##  $ luxury_assets_value     : int  22700000 8800000 33300000 23300000 29400000 13700000 29200000 1180
##  $ bank_asset_value        : int  8000000 3300000 12800000 7900000 5000000 5100000 4300000 6000000 6
##  $ loan_status             : chr  " Approved" " Rejected" " Rejected" " Rejected" ...
```

```r
# Display summary statistics
summary(loan_data)
```

```
##     loan_id     no_of_dependents  education         self_employed
##  Min.   :   1   Min.   :0.000    Length:4269        Length:4269
##  1st Qu.:1068   1st Qu.:1.000    Class :character   Class :character
##  Median :2135   Median :3.000    Mode  :character   Mode  :character
##  Mean   :2135   Mean   :2.499
##  3rd Qu.:3202   3rd Qu.:4.000
##  Max.   :4269   Max.   :5.000
##   income_annum      loan_amount         loan_term     cibil_score
##  Min.   : 200000   Min.   :  300000   Min.   : 2.0   Min.   :300.0
##  1st Qu.:2700000   1st Qu.: 7700000   1st Qu.: 6.0   1st Qu.:453.0
##  Median :5100000   Median :14500000   Median :10.0   Median :600.0
##  Mean   :5059124   Mean   :15133450   Mean   :10.9   Mean   :599.9
##  3rd Qu.:7500000   3rd Qu.:21500000   3rd Qu.:16.0   3rd Qu.:748.0
##  Max.   :9900000   Max.   :39500000   Max.   :20.0   Max.   :900.0
##  residential_assets_value commercial_assets_value luxury_assets_value
##  Min.   : -100000         Min.   :       0        Min.   :  300000
##  1st Qu.: 2200000         1st Qu.: 1300000        1st Qu.: 7500000
##  Median : 5600000         Median : 3700000        Median :14600000
##  Mean   : 7472617         Mean   : 4973155        Mean   :15126306
##  3rd Qu.:11300000         3rd Qu.: 7600000        3rd Qu.:21700000
##  Max.   :29100000         Max.   :19400000        Max.   :39200000
##  bank_asset_value   loan_status
##  Min.   :       0   Length:4269
##  1st Qu.: 2300000   Class :character
##  Median : 4600000   Mode  :character
##  Mean   : 4976692
##  3rd Qu.: 7100000
##  Max.   :14700000
```

```r
# Identify missing values
missing_values <- loan_data %>% summarise_all(~sum(is.na(.)))

# View columns with missing values
print(missing_values)
```

```
##   loan_id no_of_dependents education self_employed income_annum loan_amount
## 1       0                0         0             0            0           0
##   loan_term cibil_score residential_assets_value commercial_assets_value
## 1         0           0                        0                       0
##   luxury_assets_value bank_asset_value loan_status
## 1                   0                0           0
```

4. Data exploration and analysis

```
# number of features (attributes)
num_features <- length(colnames(loan_data))
cat("The number of features (attributes) are : ", num_features, "\n")
```

```
## The number of features (attributes) are :  13
```

```
# list the name of instances
list_features <- colnames(loan_data)
cat("\n The list of attributes are: \n")
```

```
##
##  The list of attributes are:
```

```
list_features
```

```
##  [1] "loan_id"                "no_of_dependents"
##  [3] "education"              "self_employed"
##  [5] "income_annum"           "loan_amount"
##  [7] "loan_term"              "cibil_score"
##  [9] "residential_assets_value" "commercial_assets_value"
## [11] "luxury_assets_value"    "bank_asset_value"
## [13] "loan_status"
```

5. Machine learning amethod selection and model training

6. Model evaluation

7. Model improvement/ optimization