German Credit Data Exploration_1

Dr. Prashant Mishra

3/26/2018

Outline

Main goal of this task is to predict the Class of a loan profile.

- Step 1 : Import Raw Data
- Step 2 : Clean Data
- Step 3 : Data Exploration : Visual & Statistic
- Step 4 : Prepare the data for Machine Learning Model
- Step 5 : Prepare the ML model for training and define the Cost Function
- Step 6 : Tune the parameter to minimize the cost further

1. Import Raw Data

```
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/german.data'

df <- read.table(url, sep=' ', header = 0)

head(df,n=5)
```

```
##    V1 V2  V3  V4   V5  V6  V7 V8  V9  V10 V11  V12 V13  V14  V15 V16  V17
## 1 A11  6 A34 A43 1169 A65 A75  4 A93 A101   4 A121  67 A143 A152   2 A173
## 2 A12 48 A32 A43 5951 A61 A73  2 A92 A101   2 A121  22 A143 A152   1 A173
## 3 A14 12 A34 A46 2096 A61 A74  2 A93 A101   3 A121  49 A143 A152   1 A172
## 4 A11 42 A32 A42 7882 A61 A74  2 A93 A103   4 A122  45 A143 A153   1 A173
## 5 A11 24 A33 A40 4870 A61 A73  3 A93 A101   4 A124  53 A143 A153   2 A173
##   V18  V19  V20 V21
## 1   1 A192 A201   1
## 2   1 A191 A201   2
## 3   2 A191 A201   1
## 4   2 A191 A201   1
## 5   2 A191 A201   2
```

```
str(df)
```

```
## 'data.frame':    1000 obs. of  21 variables:
##  $ V1 : Factor w/ 4 levels "A11","A12","A13",..: 1 2 4 1 1 4 4 2 4 2 ...
##  $ V2 : int  6 48 12 42 24 36 24 36 12 30 ...
##  $ V3 : Factor w/ 5 levels "A30","A31","A32",..: 5 3 5 3 4 3 3 3 3 5 ...
##  $ V4 : Factor w/ 10 levels "A40","A41","A410",..: 5 5 8 4 1 8 4 2 5 1 ...
##  $ V5 : int  1169 5951 2096 7882 4870 9055 2835 6948 3059 5234 ...
##  $ V6 : Factor w/ 5 levels "A61","A62","A63",..: 5 1 1 1 1 5 3 1 4 1 ...
##  $ V7 : Factor w/ 5 levels "A71","A72","A73",..: 5 3 4 4 3 3 5 3 4 1 ...
##  $ V8 : int  4 2 2 2 3 2 3 2 2 4 ...
```

```
##  $ V9 : Factor w/ 4 levels "A91","A92","A93",..: 3 2 3 3 3 3 3 3 1 4 ...
##  $ V10: Factor w/ 3 levels "A101","A102",..: 1 1 1 3 1 1 1 1 1 1 ...
##  $ V11: int  4 2 3 4 4 4 4 2 4 2 ...
##  $ V12: Factor w/ 4 levels "A121","A122",..: 1 1 1 2 4 4 2 3 1 3 ...
##  $ V13: int  67 22 49 45 53 35 53 35 61 28 ...
##  $ V14: Factor w/ 3 levels "A141","A142",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ V15: Factor w/ 3 levels "A151","A152",..: 2 2 2 3 3 3 2 1 2 2 ...
##  $ V16: int  2 1 1 1 2 1 1 1 1 2 ...
##  $ V17: Factor w/ 4 levels "A171","A172",..: 3 3 2 3 3 2 3 4 2 4 ...
##  $ V18: int  1 1 2 2 2 2 1 1 1 1 ...
##  $ V19: Factor w/ 2 levels "A191","A192": 2 1 1 1 1 2 1 2 1 1 ...
##  $ V20: Factor w/ 2 levels "A201","A202": 1 1 1 1 1 1 1 1 1 1 ...
##  $ V21: int  1 2 1 1 2 1 1 1 1 2 ...
```

As it is the data doesn't tell you much. We don't even know what each column represents let alone the values within the column. Our first task is to clean it so that we can perform intial exploration.

2. Clean the Raw Data

Since this dataframe is not that huge (only 21 columns and max 10 factors within each column), we will build a new DataFrame "german_credit"" from the loaded dataframe with desired details column by column.

i. Add "Class"" attribute

Adding the Class attribute to the german credit data which is the indicator of whether the profile is good or bad.

```
german_credit <- data.frame(Class = df$V21)
german_credit$Class <- 'Good'
german_credit$Class[df$V21 == 2] <- 'Bad'
```

ii. Add "CheckingAccountStatus" attribute

Adding the Class attribute to the german credit data which is the indicator of whether the profile is good or bad. -Attribute 1: (qualitative) - Status of existing checking account - A11 : ... < 0 DM (We will say 'lt.0') - A12 : 0 <= ... < 200 DM (We will say '0.to.200') - A13 : ... >= 200 DM / salary assignments for at least 1 year (We will say 'gt.200') - A14 : no checking account (We will say 'none')

```
german_credit$CheckingAccountStatus <- df$V1
levels(german_credit$CheckingAccountStatus) <- c('lt.0'
                                                 ,'0.to.200'
                                                 ,'gt.200'
                                                 ,'none'
)
```

iii. Add "Duration" attribute

- Attribute 2: (numerical)
- Duration in month

```
german_credit$Duration <- df$V2
```

iv. Add "CreditHistory" attribute

- Attribute 3: (qualitative)
- Credit history
- A30 : no credits taken/ all credits paid back duly(We will say 'NoCredit.AllPaid')
- A31 : all credits at this bank paid back duly (We will say 'ThisBank.AllPaid')

- A32 : existing credits paid back duly till now (We will say 'PaidDuly')
- A33 : delay in paying off in the past (We will say 'Delay')
- A34 : critical account/ other credits existing (not at this bank) (We will say 'Critical')

```
german_credit$CreditHistory <- df$V3
levels(german_credit$CreditHistory) <- c(
  'NoCredit.AllPaid'
  ,'ThisBank.AllPaid'
  ,'PaidDuly'
  ,'Delay'
  ,'Critical'
)
```

v. Add "Purpose" attribute

- Attribute 4: (qualitative)
- Purpose
- A40 : car (new)
- A41 : car (used)
- A42 : furniture/equipment
- A43 : radio/television
- A44 : domestic appliances
- A45 : repairs
- A46 : education
- A47 : (vacation - does not exist?)
- A48 : retraining
- A49 : business
- A410 : others

```
german_credit$Purpose <- df$V4

levels(german_credit$Purpose) <- c(
  'NewCar'
  ,'UsedCar'
  ,'Others'
  ,'Furniture.Equipment'
  ,'Radio.Television'
  ,'DomesticAppliance'
  ,'Repairs'
  ,'Education'
  ,'Retraining'
  ,'Business'
)
```

vi. Add "Amount" attribute

- Attribute 5: (numerical)
- Credit amount

```
german_credit$Amount <- df$V5
```

vii. Add "SavingsAccountBonds" attribute

- Attibute 6: (qualitative)
- Savings account/bonds
- A61 : ... < 100 DM (We will say 'lt.100')
- A62 : 100 <= ... < 500 DM (We will say '100.to.500')
- A63 : 500 <= ... < 1000 DM (We will say '500.to.1000')

- A64 : .. >= 1000 DM (We will say 'gt.1000')
- A65 : unknown/ no savings account (We will say 'Unknown')

```
german_credit$SavingsAccountBonds <- df$V6
levels(german_credit$SavingsAccountBonds) <- c(
  'lt.100'
  ,'100.to.500'
  ,'500.to.1000'
  ,'gt.1000'
  ,'Unknown'
)
```

viii. Add "EmploymentDuration" attribute

- Attribute 7: (qualitative)
- Present employment since
- A71 : unemployed (We will say 'Unemployed')
- A72 : ... < 1 year (We will say '0.to.1')
- A73 : 1 <= ... < 4 years (We will say '1.to.4')
- A74 : 4 <= ... < 7 years (We will say '4.to.7')
- A75 : .. >= 7 years (We will say 'gt.7')

```
german_credit$EmploymentDuration <- df$V7
levels(german_credit$EmploymentDuration) <- c(
  'Unemployed'
  ,'0.to.1'
  ,'1.to.4'
  ,'4.to.7'
  ,'gt.7'
)
```

ix. Add "InstallmentRatePercentage" attribute

- Attribute 8: (numerical)
- Installment rate in percentage of disposable income

```
german_credit$InstallmentRatePercentage <- df$V8
```

x. Add "Personal" attribute

- Attribute 9: (qualitative)
- Personal status and sex
- A91 : male : divorced/separated (We will say 'Male.Divorced.Seperated')
- A92 : female : divorced/separated/married (We will say 'Female.NotSingle')
- A93 : male : single (We will say 'Male.Single')
- A94 : male : married/widowed (We will say 'Male.Married.Widowed')
- A95 : female : single (We will say 'Female.Single'... does not exist)

```
german_credit$Personal <- df$V9
levels(german_credit$Personal) <- c(
  'Male.Divorced.Seperated'
  ,'Female.NotSingle'
  ,'Male.Single'
  ,'Male.Married.Widowed'
)
```

xi. Add "OtherDebtorsGuarantors" attribute

- Attribute 10: (qualitative)

- Other debtors / guarantors
- A101 : none (We will say 'None')
- A102 : co-applicant (We will say 'CoApplicant')
- A103 : guarantor (We will say 'Guarantor')

```
german_credit$OtherDebtorsGuarantors <- df$V10
levels(german_credit$OtherDebtorsGuarantors) <- c(
  'None'
  ,'CoApplicant'
  ,'Guarantor'
)
```

xii. Add "ResidenceDuration" attribute

- Attribute 11: (numerical)
- Present residence since

```
german_credit$ResidenceDuration <- df$V11
```

xiii. Add "Property" attribute

- Attribute 12: (qualitative)
- Property
- A121 : real estate (We will say 'RealEstate')
- A122 : if not A121 : building society savings agreement/ life insurance (We will say 'Insurance')
- A123 : if not A121/A122 : car or other, not in attribute 6 (We will say 'CarOther')
- A124 : unknown / no property (We will say 'Unknown')

```
german_credit$Property <- df$V12
levels(german_credit$Property) <- c(
  'RealEstate'
  ,'Insurance'
  ,'CarOther'
  ,'Unknown'
)
```

xiv. Add "Age" attribute

- Attribute 13: (numerical)
- Age in years

```
german_credit$Age <- df$V13
```

xv. Add "OtherInstallmentPlans" attribute

- Attribute 14: (qualitative)
- Other installment plans
- A141 : bank
- A142 : stores
- A143 : none

```
german_credit$OtherInstallmentPlans <- df$V14
levels(german_credit$OtherInstallmentPlans) <- c(
  'Bank'
  ,'Stores'
  ,'None'
)
```

xvi. Add "Housing" attribute

- Attribute 15: (qualitative)
- Housing
- A151 : rent
- A152 : own
- A153 : for free

```
german_credit$Housing <- df$V15
levels(german_credit$Housing) <- c('Rent', 'Own', 'ForFree')
```

xvii. Add "NumberExistingCredits" attribute

- Attribute 16: (numerical)
- Number of existing credits at this bank

```
german_credit$NumberExistingCredits <- df$V16
```

xviii. Add "Job" attribute

- Attribute 17: (qualitative)
- Job
- A171 : unemployed/ unskilled - non-resident (We will say 'UnemployedUnskilled')
- A172 : unskilled - resident (We will say 'UnskilledResident')
- A173 : skilled employee / official (We will say 'SkilledEmployee')
- A174 : management/ self-employed/ (We will say 'Management.SelfEmp.HighlyQualified') highly qualified employee/ officer

```
german_credit$Job <- df$V17
levels(german_credit$Job) <- c(
  'UnemployedUnskilled'
  ,'UnskilledResident'
  ,'SkilledEmployee'
  ,'Management.SelfEmp.HighlyQualified'
)
```

xix. Add "NumberPeopleMaintenance" attribute

- Attribute 18: (numerical)
- Number of people being liable to provide maintenance for

```
german_credit$NumberPeopleMaintenance <- df$V18
```

xx. Add "Telephone" attribute

- Attribute 19: (qualitative)
- Telephone
- A191 : none (We will say 0)
- A192 : yes, registered under the customers name (We will say 1)

```
german_credit$Telephone <- df$V19
levels(german_credit$Telephone) <- c(
  0
  ,1
)
```

xxi. Add "ForeignWorker" attribute

- Attribute 20: (qualitative)
- foreign worker
- A201 : yes
- A202 : no

```
german_credit$ForeignWorker <- df$V20
levels(german_credit$ForeignWorker) <- c(1, 0)
```

Finally let's save all these changes into a file, which we can import anytime without going through all these steps.

```
save(german_credit, file = 'german_credit')
write.csv(german_credit, 'german_credit_full.csv',
          row.names = FALSE)
```