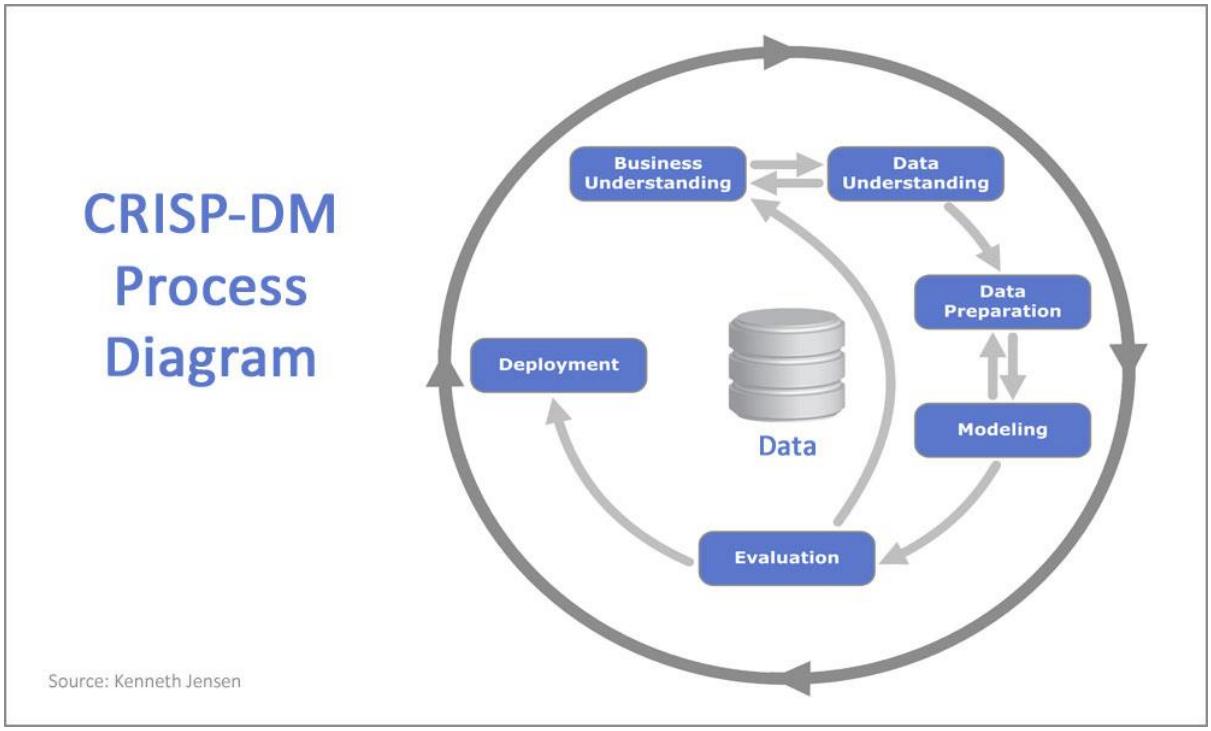
CREDIT RISK CLASSIFICATION



1. **Bussiness Understanding**

Karena banyaknya nasabah ingin mengajukan pinjaman terhadap bank, bank mengkategorikan nasabah berdasarkan jenis *Credit Risk*. *Credit Risk* adalah kemungkinan kerugian yang diakibatkan oleh kegagalan peminjam untuk membayar kembali pinjaman atau memenuhi kewajiban kontraktual. *Credit Risk* dihitung berdasarkan kemampuan nasabah secara keseluruhan untuk membayar kembali pinjaman sesuai dengan nominal pinjaman. *Credit Risk* dikategorikan menjadi “bagus” dan “ tidak bagus” . Jika *credit risk* tergolong bagus, maka risiko nasabah gagal bayar tergolong rendah . Sebaliknya, jika credis risk tergolong “tidak bagus” , maka risiko nasabah gagal bayar tergolong tinggi.

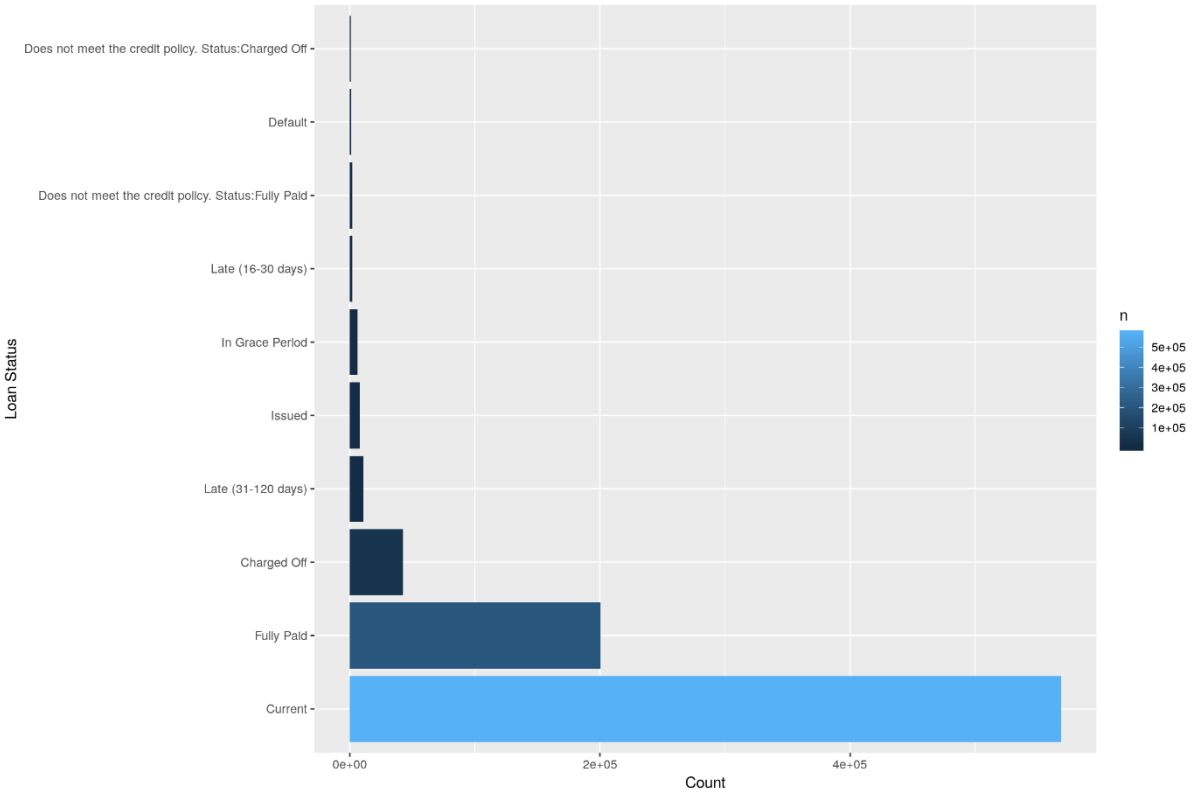
Untuk mempermudah pengklasifikasian *credit risk* oleh bank, maka dibuat mesin yang dapat mengklasifikasikan *credit risk* nasabah berdasarkan variabel tertentu. Tujuannya agar klasifikasi *credit risk* menjadi mudah dan efisien dibandingkan dengan cara manual.

2. Data Understanding

Variabel indepen yang digunakan yaitu *Loan Amount, Interest Rate, Grade , employment length, home ownership, Total annual income,* dan *Payment Term.* Sedangkan variabel dependen yaitu *Loan Status.*

```
## # A tibble: 887,379 x 8
##    loan_status loan_amnt int_rate grade emp_length home_ownership
##    <chr>           <dbl>    <dbl> <chr> <chr>      <chr>
##  1 Fully Paid       5000     10.6 B     10+ years  RENT
##  2 Charged Off      2500     15.3 C     < 1 year   RENT
##  3 Fully Paid       2400     16.0 C     10+ years  RENT
##  4 Fully Paid      10000     13.5 C     10+ years  RENT
##  5 Current          3000     12.7 B     1 year     RENT
##  6 Fully Paid       5000      7.9 A     3 years    RENT
##  7 Current          7000     16.0 C     8 years    RENT
##  8 Fully Paid       3000     18.6 E     9 years    RENT
##  9 Charged Off      5600     21.3 F     4 years    OWN
## 10 Charged Off      5375     12.7 B     < 1 year   RENT
## # ... with 887,369 more rows, and 2 more variables: annual_inc <dbl>,
## #   term <chr>
```

Deskripsi diatas menunjukkan jumlah data 887.879 , 8 variabel, dan isi data.



Grafik di atas menunjukan frekuensi per kategori pada *Loan Status*. Kategori *Current* memiliki frekuensi yang paling tinggi kemudian diikuti dengan Full Paid dan seterusnya.

Grafik di atas adalah pengujian apakah *p*endapatan tahunan berpengaruh terhadap besar pinjaman. Berdasarkan grafik diatas, semakin besar pendapatan tahunan nasabah semakin besar pula nominal peminjaman nasabah.

3. Data Preparation

Di tahap ini, data akan dilakukan cleansing missing value pada variabel *annual income, home ownership* dan *employment length*.

```
# Remove the 4 rows with missing annual income, 49 rows where home ownership is 'NONE' or 'AN
Y' and rows where emp_length is 'n/a'.

loan = loan %>%
        filter(!is.na(annual_inc) ,
               !(home_ownership %in% c('NONE' , 'ANY')) ,
               emp_length != 'n/a')
```

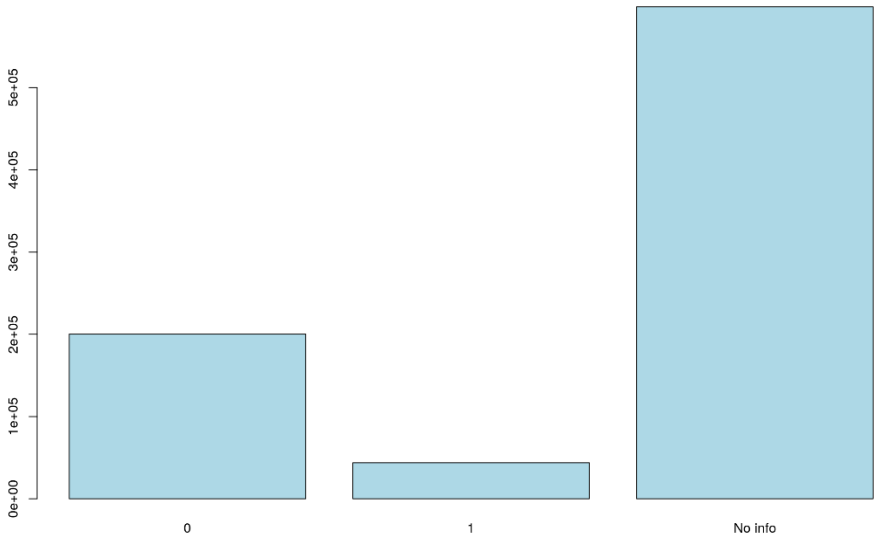
Agar memudahkan perhitungan, kami ingin mengonversi variabel ini ke biner (1 untuk default dan 0 untuk non-default) tetapi kami memiliki 10 level berbeda. Pinjaman dengan status status *Current, Late payments, In grace period* perlu dihapus. Oleh karena itu, kami membuat variabel baru yang disebut *loan_outcome* dimana maka vaiabel *loan status*, diubah menjadi biner dengan keterangan

loan_outcome -> 1 if loan_status = ‘Charged Off’ or ‘Default’ loan_outcome -> 0

if loan_status = ‘Fully Paid’

```
loan = loan %>%
        mutate(loan_outcome = ifelse(loan_status %in% c('Charged Off' , 'Default') ,
                                     1,
                                     ifelse(loan_status == 'Fully Paid' , 0 , 'No info')
                                     ))

barplot(table(loan$loan_outcome) , col = 'lightblue')
```



Kami akan membuat dataset baru yang hanya berisi baris dengan 0 atau 1 dalam fitur loan_outcome untuk pemodelan yang lebih baik.

```
# Create the new dataset by filtering 0's and 1's in the loan_outcome column and remove loan_s
tatus column for the modelling
loan2 = loan %>%
        select(-loan_status) %>%
        filter(loan_outcome %in% c(0 , 1))
```

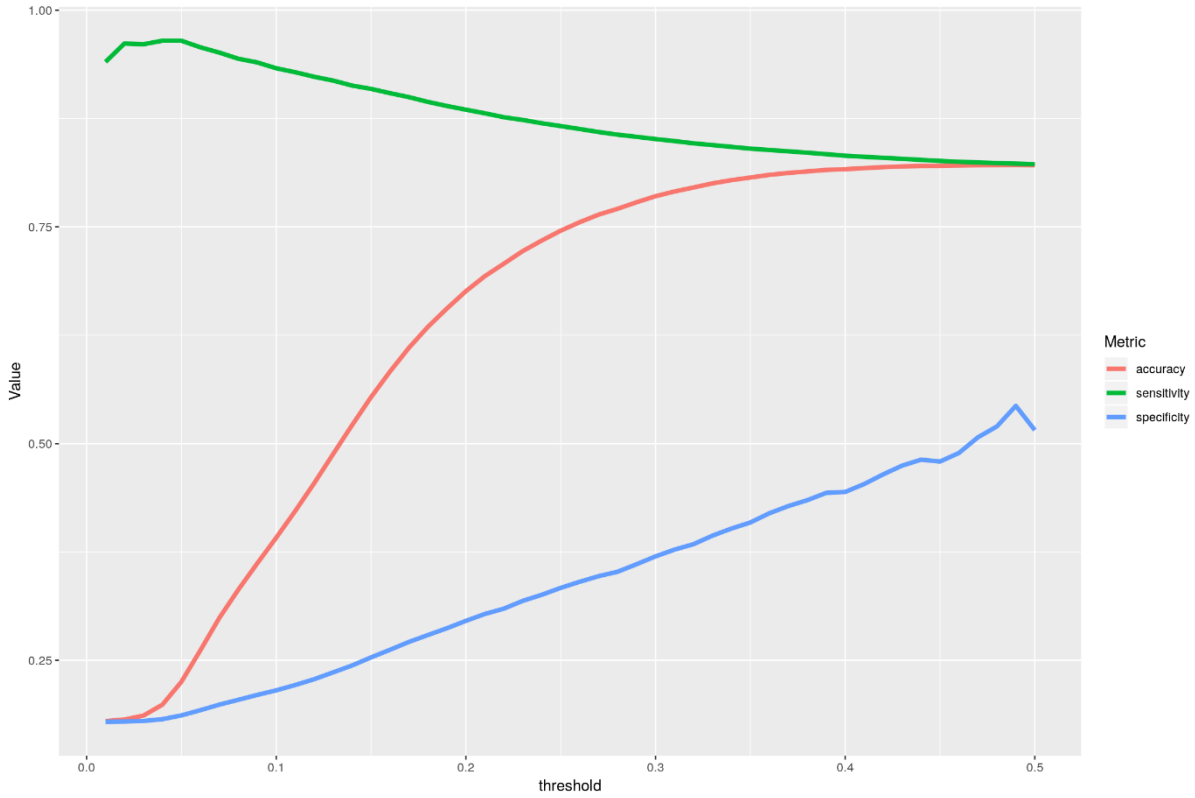
Jumlah dataset yang baru yaitu **244.179.**

4. Modelling

Proses Pemodelan:

a) Kami membuat binary loan_outcome yang akan menjadi variabel respons kami.
b) Kami mengecualikan beberapa variabel independen untuk membuat model lebih sederhana.
c) Kami membagi dataset menjadi training set (75%) dan testing set (25%) untuk validasi.
d) Kami melatih model untuk memprediksi probabilitas default.
e) Model yang digunakan yaitu *Logistic Regression*

5. Evaluation

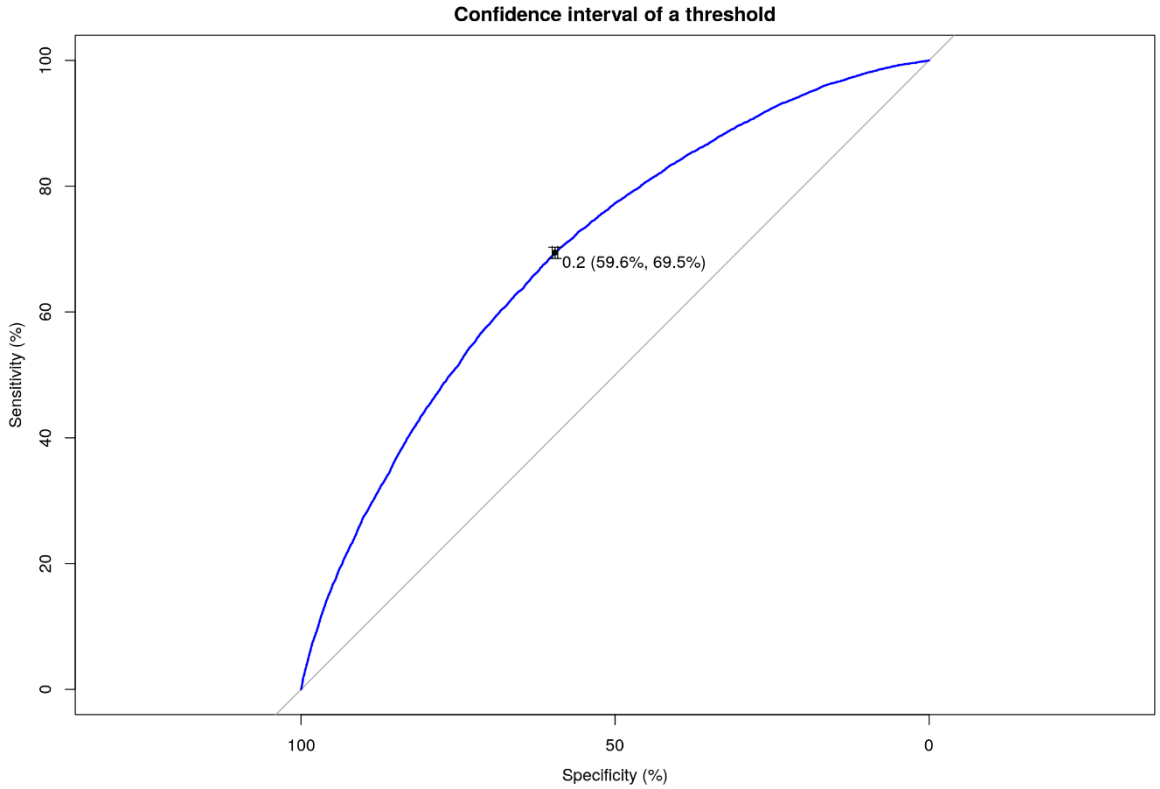


Threshold 25% - 30% tampaknya ideal karena peningkatan persentase pemotongan lebih lanjut tidak berdampak signifikan terhadap keakuratan model. Untuk *Confusion Matrix* memiliki titik potong 30% adalah ini,

```
##          Actual
## Predicted     0     1
##         0 44853  7834
##         1  5266  3092
```

```
## [1] "Accuracy : 0.7854"
```

Kurva ROC (*Receiver Operating Characteristics*)



Model *Logistic Regression* digunakan untuk memprediksi status pinjaman. *Cut off* yang berbeda digunakan untuk memutuskan apakah pinjaman harus diberikan atau tidak. *Cut off* 30% memberikan akurasi yang baik sebesar 78,54%. Keputusan untuk menetapkan *cut off* adalah sewenang-wenang dan tingkat *threshold* yang lebih tinggi meningkatkan risiko. Area *Under Curve* juga memberikan ukuran akurasi, yang menjadi 69,57%.

6. Deployment

Pada proses *deployment*, model ini dilakukan proses *deployment* kedalam website sehingga dapat diakses secara online dan tidak membebani kinerja computer.