

BUAN4310 Group Project 1

Shayla Nguyen

2024-10-25

```
library(caret)

## Loading required package: ggplot2
## Loading required package: lattice

library(ROSE)

## Loaded ROSE 0.0-4

library(pROC)

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

library(readr)

# Load the credit31 data
credit31 <- read_csv("credit_31.csv")

## New names:
## • `` -> `...1`

## Rows: 30000 Columns: 68
## — Column specification —————
## Delimiter: ","
## chr (12): NAME_CONTRACT_TYPE, CODE_GENDER, FLAG_OWN_CAR, FLAG_OWN_REALTY,
NA...
## dbl (56): ...1, SK_ID_CURR, TARGET, CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CR
ED...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this m
essage.

# Add new fields into data frame to improve model accuracy
credit31$Income_Credit_Ratio <- credit31$AMT_INCOME_TOTAL / credit31$AMT_CRED
```

```
IT
credit31$Annuity_Income_Ratio <- credit31$AMT_ANNUITY / credit31$AMT_INCOME_T
OTAL
credit31$Credit_As_Percentage <- credit31$AMT_CREDIT / credit31$AMT_INCOME_TO
TAL
credit31$Percent_Days_Employed <- credit31$DAYS_EMPLOYED / credit31$DAYS_BIRT
H
credit31$Income_Per_Person <- credit31$AMT_INCOME_TOTAL / credit31$CNT_FAM_ME
MBERS
```

Remove XNA from CODE_GENDER variable and convert to factor

```
credit31 <- credit31[credit31$CODE_GENDER != "XNA", ]
credit31$CODE_GENDER <- factor(credit31$CODE_GENDER)
```

Explore data

```
names(credit31)
```

```
## [1] "...1" "SK_ID_CURR"
## [3] "TARGET" "NAME_CONTRACT_TYPE"
## [5] "CODE_GENDER" "FLAG_OWN_CAR"
## [7] "FLAG_OWN_REALTY" "CNT_CHILDREN"
## [9] "AMT_INCOME_TOTAL" "AMT_CREDIT"
## [11] "AMT_ANNUITY" "AMT_GOODS_PRICE"
## [13] "NAME_TYPE_SUITE" "NAME_INCOME_TYPE"
## [15] "NAME_EDUCATION_TYPE" "NAME_FAMILY_STATUS"
## [17] "NAME_HOUSING_TYPE" "DAYS_BIRTH"
## [19] "DAYS_EMPLOYED" "DAYS_REGISTRATION"
## [21] "DAYS_ID_PUBLISH" "OWN_CAR_AGE"
## [23] "FLAG_MOBIL" "FLAG_EMP_PHONE"
## [25] "FLAG_WORK_PHONE" "FLAG_CONT_MOBILE"
## [27] "FLAG_PHONE" "FLAG_EMAIL"
## [29] "OCCUPATION_TYPE" "CNT_FAM_MEMBERS"
## [31] "REGION_RATING_CLIENT" "REGION_RATING_CLIENT_W_CITY"
## [33] "WEEKDAY_APPR_PROCESS_START" "HOUR_APPR_PROCESS_START"
## [35] "REG_REGION_NOT_LIVE_REGION" "REG_REGION_NOT_WORK_REGION"
## [37] "LIVE_REGION_NOT_WORK_REGION" "REG_CITY_NOT_LIVE_CITY"
## [39] "REG_CITY_NOT_WORK_CITY" "LIVE_CITY_NOT_WORK_CITY"
## [41] "ORGANIZATION_TYPE" "DAYS_LAST_PHONE_CHANGE"
## [43] "FLAG_DOCUMENT_2" "FLAG_DOCUMENT_3"
## [45] "FLAG_DOCUMENT_4" "FLAG_DOCUMENT_5"
## [47] "FLAG_DOCUMENT_6" "FLAG_DOCUMENT_7"
## [49] "FLAG_DOCUMENT_8" "FLAG_DOCUMENT_9"
## [51] "FLAG_DOCUMENT_10" "FLAG_DOCUMENT_11"
## [53] "FLAG_DOCUMENT_12" "FLAG_DOCUMENT_13"
## [55] "FLAG_DOCUMENT_14" "FLAG_DOCUMENT_15"
## [57] "FLAG_DOCUMENT_16" "FLAG_DOCUMENT_17"
## [59] "FLAG_DOCUMENT_18" "FLAG_DOCUMENT_19"
## [61] "FLAG_DOCUMENT_20" "FLAG_DOCUMENT_21"
## [63] "AMT_REQ_CREDIT_BUREAU_HOUR" "AMT_REQ_CREDIT_BUREAU_DAY"
## [65] "AMT_REQ_CREDIT_BUREAU_WEEK" "AMT_REQ_CREDIT_BUREAU_MON"
```

```
## [67] "AMT_REQ_CREDIT_BUREAU_QRT"    "AMT_REQ_CREDIT_BUREAU_YEAR"
## [69] "Income_Credit_Ratio"          "Annuity_Income_Ratio"
## [71] "Credit_As_Percentage"         "Percent_Days_Employed"
## [73] "Income_Per_Person"
```

```
str(credit31)
```

```
## tibble [30,000 × 73] (S3: tbl_df/tbl/data.frame)
## $ ...1 : num [1:30000] 284834 161354 132607 199508
99768 ...
## $ SK_ID_CURR : num [1:30000] 429876 287055 253803 331291
215821 ...
## $ TARGET : num [1:30000] 0 0 1 0 0 0 0 0 0 ...
## $ NAME_CONTRACT_TYPE : chr [1:30000] "Cash loans" "Cash loans" "C
ash loans" "Cash loans" ...
## $ CODE_GENDER : Factor w/ 2 levels "F","M": 1 2 2 1 1 2 1
1 1 1 ...
## $ FLAG_OWN_CAR : chr [1:30000] "N" "Y" "N" "Y" ...
## $ FLAG_OWN_REALTY : chr [1:30000] "N" "Y" "Y" "N" ...
## $ CNT_CHILDREN : num [1:30000] 0 0 0 0 0 0 0 0 0 ...
## $ AMT_INCOME_TOTAL : num [1:30000] 103500 81000 112500 225000 6
7500 ...
## $ AMT_CREDIT : num [1:30000] 675000 808650 423000 646920
135000 ...
## $ AMT_ANNUITY : num [1:30000] 21776 26217 28269 25065 1615
0 ...
## $ AMT_GOODS_PRICE : num [1:30000] 675000 675000 423000 540000
135000 ...
## $ NAME_TYPE_SUITE : chr [1:30000] "Family" "Family" "Spouse, p
artner" "Unaccompanied" ...
## $ NAME_INCOME_TYPE : chr [1:30000] "State servant" "Working" "W
orking" "Working" ...
## $ NAME_EDUCATION_TYPE : chr [1:30000] "Higher education" "Secondar
y / secondary special" "Secondary / secondary special" "Higher education" ...
## $ NAME_FAMILY_STATUS : chr [1:30000] "Separated" "Married" "Marri
ed" "Married" ...
## $ NAME_HOUSING_TYPE : chr [1:30000] "House / apartment" "House /
apartment" "House / apartment" "House / apartment" ...
## $ DAYS_BIRTH : num [1:30000] -14211 -17884 -14629 -12894
-16825 ...
## $ DAYS_EMPLOYED : num [1:30000] -2875 -2192 -984 -1994 -1087
...
## $ DAYS_REGISTRATION : num [1:30000] -4018 -7442 -741 -1278 -8220
...
## $ DAYS_ID_PUBLISH : num [1:30000] -4693 -1428 -1747 -3897 -367
...
## $ OWN_CAR_AGE : num [1:30000] NA 1 NA 1 18 NA NA NA NA NA
...
## $ FLAG_MOBIL : num [1:30000] 1 1 1 1 1 1 1 1 1 1 ...
## $ FLAG_EMP_PHONE : num [1:30000] 1 1 1 1 1 1 1 0 1 0 ...
```

```

## $ FLAG_WORK_PHONE           : num [1:30000] 1 1 0 1 0 0 0 0 0 0 ...
## $ FLAG_CONT_MOBILE         : num [1:30000] 1 1 1 1 1 1 1 1 1 1 ...
## $ FLAG_PHONE                : num [1:30000] 0 1 0 1 0 0 0 0 0 0 ...
## $ FLAG_EMAIL                : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ OCCUPATION_TYPE           : chr [1:30000] "Laborers" "Security staff"
"Laborers" "Laborers" ...
## $ CNT_FAM_MEMBERS           : num [1:30000] 1 2 2 2 2 1 2 2 1 2 ...
## $ REGION_RATING_CLIENT      : num [1:30000] 2 2 2 2 2 2 2 2 2 3 ...
## $ REGION_RATING_CLIENT_W_CITY : num [1:30000] 2 2 2 2 2 2 2 2 2 3 ...
## $ WEEKDAY_APPR_PROCESS_START : chr [1:30000] "SUNDAY" "TUESDAY" "MONDAY"
"SATURDAY" ...
## $ HOUR_APPR_PROCESS_START    : num [1:30000] 11 11 8 10 9 12 14 11 11 11
...
## $ REG_REGION_NOT_LIVE_REGION : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ REG_REGION_NOT_WORK_REGION : num [1:30000] 0 0 1 0 0 0 0 0 0 0 ...
## $ LIVE_REGION_NOT_WORK_REGION : num [1:30000] 0 0 1 0 0 0 0 0 0 0 ...
## $ REG_CITY_NOT_LIVE_CITY     : num [1:30000] 0 0 0 0 0 1 0 0 0 0 ...
## $ REG_CITY_NOT_WORK_CITY     : num [1:30000] 0 1 1 0 1 1 1 0 0 0 ...
## $ LIVE_CITY_NOT_WORK_CITY    : num [1:30000] 0 1 1 0 1 0 1 0 0 0 ...
## $ ORGANIZATION_TYPE          : chr [1:30000] "Postal" "Business Entity Ty
pe 3" "Industry: type 9" "Business Entity Type 1" ...
## $ DAYS_LAST_PHONE_CHANGE     : num [1:30000] -1735 0 -570 -1748 -1204 ...
## $ FLAG_DOCUMENT_2            : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ FLAG_DOCUMENT_3            : num [1:30000] 1 1 1 1 1 1 1 0 0 0 ...
## $ FLAG_DOCUMENT_4            : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ FLAG_DOCUMENT_5            : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ FLAG_DOCUMENT_6            : num [1:30000] 0 0 0 0 0 0 0 0 0 1 ...
## $ FLAG_DOCUMENT_7            : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ FLAG_DOCUMENT_8            : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ FLAG_DOCUMENT_9            : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ FLAG_DOCUMENT_10           : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ FLAG_DOCUMENT_11           : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ FLAG_DOCUMENT_12           : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ FLAG_DOCUMENT_13           : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ FLAG_DOCUMENT_14           : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ FLAG_DOCUMENT_15           : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ FLAG_DOCUMENT_16           : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ FLAG_DOCUMENT_17           : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ FLAG_DOCUMENT_18           : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ FLAG_DOCUMENT_19           : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ FLAG_DOCUMENT_20           : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ FLAG_DOCUMENT_21           : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ AMT_REQ_CREDIT_BUREAU_HOUR : num [1:30000] 0 0 0 1 0 0 0 0 0 0 ...
## $ AMT_REQ_CREDIT_BUREAU_DAY  : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ AMT_REQ_CREDIT_BUREAU_WEEK : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ AMT_REQ_CREDIT_BUREAU_MON  : num [1:30000] 0 0 0 1 0 0 0 0 1 1 ...
## $ AMT_REQ_CREDIT_BUREAU_QRT  : num [1:30000] 0 0 0 0 0 0 0 0 0 0 ...
## $ AMT_REQ_CREDIT_BUREAU_YEAR : num [1:30000] 2 2 0 2 1 1 2 5 0 2 ...
## $ Income_Credit_Ratio        : num [1:30000] 0.153 0.1 0.266 0.348 0.5 ..
.

```

```
## $ Annuity_Income_Ratio      : num [1:30000] 0.21 0.324 0.251 0.111 0.239
...
## $ Credit_As_Percentage      : num [1:30000] 6.52 9.98 3.76 2.88 2 ...
## $ Percent_Days_Employed     : num [1:30000] 0.2023 0.1226 0.0673 0.1546
0.0646 ...
## $ Income_Per_Person         : num [1:30000] 103500 40500 56250 112500 33
750 ...
```

```
summary(credit31$TARGET)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.0000  0.0000  0.0000  0.1949  0.0000  1.0000
```

```
# Convert education type to factor with levels across education
```

```
credit31$NAME_EDUCATION_TYPE <- factor(credit31$NAME_EDUCATION_TYPE, levels =
c(
  "Secondary / secondary special",
  "Higher education",
  "Lower secondary",
  "Incomplete higher",
  "Academic degree"))
```

```
# Set Target variable as factor
```

```
credit31$TARGET <- as.factor(credit31$TARGET)
```

```
# Variable List
```

```
# Percent_Days_Employed, NAME_EDUCATION_TYPE, REGION_RATING_CLIENT_W_CITY, AM
T_GOODS_PRICE, CODE_GENDER, DAYS_BIRTH, AMT_CREDIT, AMT_ANNUIITY, DAYS_EMPLOYE
D, DAYS_REGISTRATION, DAYS_ID_PUBLISH, Annuity_Income_Ratio
```

```
# Remove unused variables
```

```
credit31 <- credit31[ , -c(1:2, 4, 6:9, 13:14, 16:17, 22:31, 33:69, 71, 73)]
names(credit31)
```

```
## [1] "TARGET"                                "CODE_GENDER"
## [3] "AMT_CREDIT"                            "AMT_ANNUIITY"
## [5] "AMT_GOODS_PRICE"                       "NAME_EDUCATION_TYPE"
## [7] "DAYS_BIRTH"                            "DAYS_EMPLOYED"
## [9] "DAYS_REGISTRATION"                     "DAYS_ID_PUBLISH"
## [11] "REGION_RATING_CLIENT_W_CITY"           "Annuity_Income_Ratio"
## [13] "Percent_Days_Employed"
```

```
# Training - Validation split
```

```
set.seed(666)
```

```
train_index <- sample(1:nrow(credit31), 0.7 * nrow(credit31))
```

```
valid_index <- setdiff(1:nrow(credit31), train_index)
```

```
train_df <- credit31[train_index, ]
```

```
valid_df <- credit31[valid_index, ]
```

```
# Double check
```

```
nrow(train_df)
```

```
## [1] 21000

nrow(valid_df)

## [1] 9000

head(train_df)

## # A tibble: 6 × 13
##   TARGET CODE_GENDER AMT_CREDIT AMT_ANNUIITY AMT_GOODS_PRICE NAME_EDUCATION
##   _TYPE
##   <fct> <fct>          <dbl>      <dbl>          <dbl> <fct>
## 1 0      M            956574      38066.          855000 Secondary / se
conda...
## 2 0      F            1633473      45050.          1363500 Secondary / se
conda...
## 3 1      F            279000      15134.          279000 Secondary / se
conda...
## 4 0      F            405000      20250          405000 Higher educati
on
## 5 1      M            279000      22041          279000 Secondary / se
conda...
## 6 0      F            808650      26217          675000 Secondary / se
conda...
## # i 7 more variables: DAYS_BIRTH <dbl>, DAYS_EMPLOYED <dbl>,
## #   DAYS_REGISTRATION <dbl>, DAYS_ID_PUBLISH <dbl>,
## #   REGION_RATING_CLIENT_W_CITY <dbl>, Annuity_Income_Ratio <dbl>,
## #   Percent_Days_Employed <dbl>

head(valid_df)

## # A tibble: 6 × 13
##   TARGET CODE_GENDER AMT_CREDIT AMT_ANNUIITY AMT_GOODS_PRICE NAME_EDUCATION
##   _TYPE
##   <fct> <fct>          <dbl>      <dbl>          <dbl> <fct>
## 1 0      M            808650      26217          675000 Secondary / se
conda...
## 2 1      M            423000      28269          423000 Secondary / se
conda...
## 3 0      M            450000      30074.          450000 Secondary / se
conda...
## 4 0      F            202500      10125          202500 Secondary / se
conda...
## 5 0      F            269550      12002.          225000 Secondary / se
conda...
## 6 0      F            1125000      36292.          1125000 Secondary / se
conda...
## # i 7 more variables: DAYS_BIRTH <dbl>, DAYS_EMPLOYED <dbl>,
## #   DAYS_REGISTRATION <dbl>, DAYS_ID_PUBLISH <dbl>,
## #   REGION_RATING_CLIENT_W_CITY <dbl>, Annuity_Income_Ratio <dbl>,
## #   Percent_Days_Employed <dbl>
```

```
str(train_df)
```

```
## tibble [21,000 × 13] (S3: tbl_df/tbl/data.frame)
## $ TARGET                                     : Factor w/ 2 levels "0","1": 1 1 2 1 2 1 1
1 1 1 ...
## $ CODE_GENDER                               : Factor w/ 2 levels "F","M": 2 1 1 1 2 1 2
2 1 2 ...
## $ AMT_CREDIT                                : num [1:21000] 956574 1633473 279000 405000
279000 ...
## $ AMT_ANNUITY                               : num [1:21000] 38066 45050 15134 20250 2204
1 ...
## $ AMT_GOODS_PRICE                           : num [1:21000] 855000 1363500 279000 405000
279000 ...
## $ NAME_EDUCATION_TYPE                       : Factor w/ 5 levels "Secondary / secondary
special",...: 1 1 1 2 1 1 1 2 1 1 ...
## $ DAYS_BIRTH                                : num [1:21000] -14523 -13597 -21630 -14352
-23121 ...
## $ DAYS_EMPLOYED                             : num [1:21000] -926 -247 -14068 -4132 36524
3 ...
## $ DAYS_REGISTRATION                         : num [1:21000] -8452 -7709 -5517 -10 -935 .
..
## $ DAYS_ID_PUBLISH                           : num [1:21000] -4476 -4795 -5024 -2199 -445
0 ...
## $ REGION_RATING_CLIENT_W_CITY: num [1:21000] 2 3 2 2 2 2 2 2 2 2 ...
## $ Annuity_Income_Ratio                     : num [1:21000] 0.169 0.222 0.108 0.15 0.109
...
## $ Percent_Days_Employed                     : num [1:21000] 0.0638 0.0182 0.6504 0.2879
-15.797 ...
```

```
str(valid_df)
```

```
## tibble [9,000 × 13] (S3: tbl_df/tbl/data.frame)
## $ TARGET                                     : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 2
1 1 1 ...
## $ CODE_GENDER                               : Factor w/ 2 levels "F","M": 2 2 2 1 1 1 1
1 1 2 ...
## $ AMT_CREDIT                                : num [1:9000] 808650 423000 450000 202500 2
69550 ...
## $ AMT_ANNUITY                               : num [1:9000] 26217 28269 30074 10125 12002
...
## $ AMT_GOODS_PRICE                           : num [1:9000] 675000 423000 450000 202500 2
25000 ...
## $ NAME_EDUCATION_TYPE                       : Factor w/ 5 levels "Secondary / secondary
special",...: 1 1 1 1 1 1 4 1 1 1 ...
## $ DAYS_BIRTH                                : num [1:9000] -17884 -14629 -9655 -21512 -2
2485 ...
## $ DAYS_EMPLOYED                             : num [1:9000] -2192 -984 -2940 -1874 365243
...
## $ DAYS_REGISTRATION                         : num [1:9000] -7442 -741 -8153 -10778 -1454
4 ...
```

```

## $ DAYS_ID_PUBLISH          : num [1:9000] -1428 -1747 -2298 -4811 -4620
...
## $ REGION_RATING_CLIENT_W_CITY: num [1:9000] 2 2 2 2 3 2 2 2 1 2 ...
## $ Annuity_Income_Ratio      : num [1:9000] 0.324 0.251 0.122 0.113 0.133
...
## $ Percent_Days_Employed     : num [1:9000] 0.1226 0.0673 0.3045 0.0871 -
16.2439 ...

# Use ROSE to balance model
train_df_rose <- ROSE(TARGET ~ Percent_Days_Employed + NAME_EDUCATION_TYPE +
REGION_RATING_CLIENT_W_CITY + AMT_GOODS_PRICE + CODE_GENDER + DAYS_BIRTH + AM
T_CREDIT + AMT_ANNUITY + DAYS_EMPLOYED + DAYS_REGISTRATION + DAYS_ID_PUBLISH
+ Annuity_Income_Ratio,
                      data = train_df, seed = 666)$data

table(train_df_rose$TARGET)

##
##      0      1
## 10339 10642

# Normalization algorithm
train_norm <- train_df_rose
valid_norm <- valid_df

norm_values <- preProcess(train_df_rose[, -c(1)],
                          method = c("center", "scale"))
train_norm[, -c(1)] <- predict(norm_values,
                              train_df_rose[, -c(1)])

# Apply to validation set
valid_norm[, -c(1)] <- predict(norm_values,
                              valid_df[, -c(1)])

# Drop missing values
library(tidyr)
valid_norm <- drop_na(valid_norm)

# Train logistic regression model
logistic_model <- glm(TARGET ~ Percent_Days_Employed + NAME_EDUCATION_TYPE +
REGION_RATING_CLIENT_W_CITY + AMT_GOODS_PRICE + CODE_GENDER + DAYS_BIRTH + AM
T_CREDIT + AMT_ANNUITY + DAYS_EMPLOYED + DAYS_REGISTRATION + DAYS_ID_PUBLISH
+ Annuity_Income_Ratio,
                      data = train_norm, family = binomial)

# Prediction on training set
logistic_pred_train <- predict(logistic_model, newdata = train_norm, type = "
response")
logistic_pred_train_class <- ifelse(logistic_pred_train > 0.5, 1, 0)

```



```

# Prediction on validation set
logistic_pred_valid <- predict(logistic_model, newdata = valid_norm, type = "
response")
logistic_pred_valid_class <- ifelse(logistic_pred_valid > 0.5, 1, 0)

# Confusion matrix on training set
confusionMatrix(as.factor(logistic_pred_train_class), as.factor(train_norm$TA
RGET), positive = "1")

## Confusion Matrix and Statistics
##
##              Reference
## Prediction    0    1
##              0 5773 3900
##              1 4566 6742
##
##              Accuracy : 0.5965
##              95% CI : (0.5898, 0.6031)
##      No Information Rate : 0.5072
##      P-Value [Acc > NIR] : < 2.2e-16
##
##              Kappa : 0.1921
##
##  Mcnemar's Test P-Value : 4.923e-13
##
##              Sensitivity : 0.6335
##              Specificity : 0.5584
##              Pos Pred Value : 0.5962
##              Neg Pred Value : 0.5968
##              Prevalence : 0.5072
##              Detection Rate : 0.3213
##      Detection Prevalence : 0.5390
##              Balanced Accuracy : 0.5959
##
##              'Positive' Class : 1
##

# Confusion matrix on validation set
confusionMatrix(as.factor(logistic_pred_valid_class), as.factor(valid_norm$TA
RGET), positive = "1")

## Confusion Matrix and Statistics
##
##              Reference
## Prediction    0    1
##              0 4027  600
##              1 3198 1165
##
##              Accuracy : 0.5775
##              95% CI : (0.5672, 0.5878)

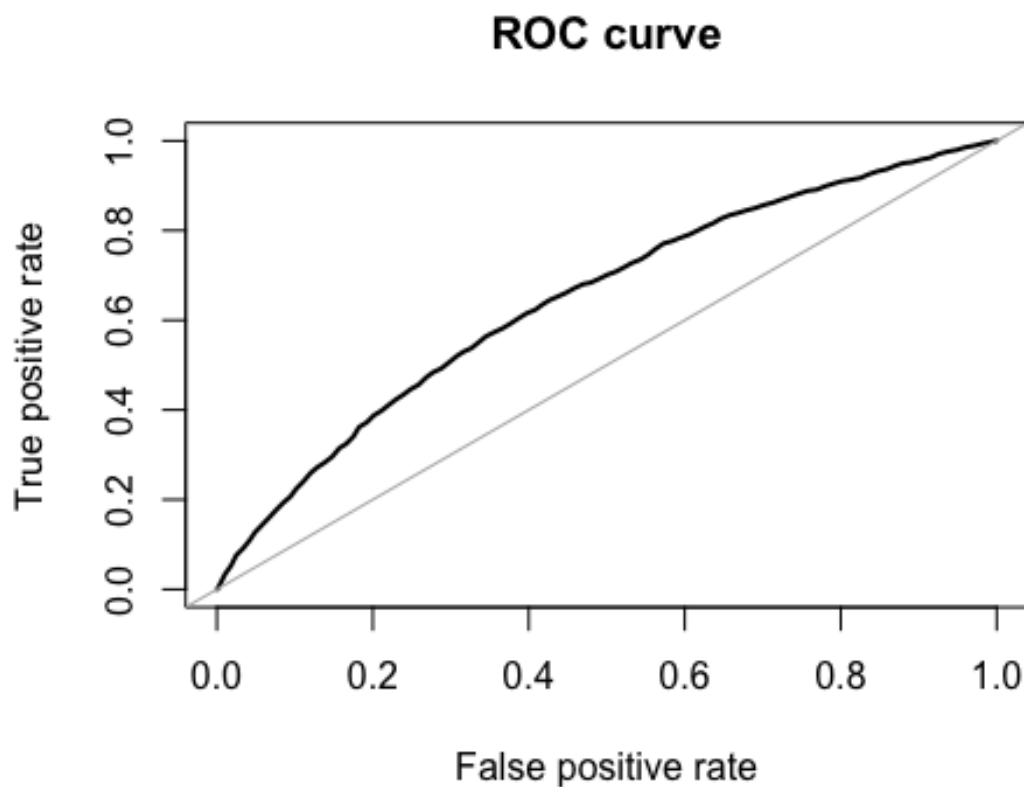
```

```
##      No Information Rate : 0.8037
##      P-Value [Acc > NIR] : 1
##
##      Kappa : 0.1397
##
##      Mcnemar's Test P-Value : <2e-16
##
##      Sensitivity : 0.6601
##      Specificity : 0.5574
##      Pos Pred Value : 0.2670
##      Neg Pred Value : 0.8703
##      Prevalence : 0.1963
##      Detection Rate : 0.1296
##      Detection Prevalence : 0.4853
##      Balanced Accuracy : 0.6087
##
##      'Positive' Class : 1
##
```

Model Evaluation

```
library(ROSE)
```

```
ROSE::roc.curve(valid_norm$TARGET, logistic_pred_valid)
```



```

## Area under the curve (AUC): 0.646

# Load new customer data
new_customers <- read_csv("credit_test_31.csv")

## New names:
## Rows: 5 Columns: 67
## — Column specification
## _____ Delimiter: "," chr
r
## (12): NAME_CONTRACT_TYPE, CODE_GENDER, FLAG_OWN_CAR, FLAG_OWN_REALTY, NA..
. dbl
## (55): ...1, SK_ID_CURR, CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CREDIT, AMT..
.
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this mes
sage.
## • `` -> `...1`

# Preprocess new customer data
new_customers$Income_Credit_Ratio <- new_customers$AMT_INCOME_TOTAL / new_cus
tomers$AMT_CREDIT
new_customers$Annuity_Income_Ratio <- new_customers$AMT_ANNUITY / new_custome
rs$AMT_INCOME_TOTAL
new_customers$Credit_As_Percentage <- new_customers$AMT_CREDIT / new_customer
s$AMT_INCOME_TOTAL
new_customers$Percent_Days_Employed <- new_customers$DAYS_EMPLOYED / new_cust
omers$DAYS_BIRTH
new_customers$Income_Per_Person <- new_customers$AMT_INCOME_TOTAL / new_custo
mers$CNT_FAM_MEMBERS

# Remove XNA from CODE_GENDER variable and convert to factor
new_customers <- new_customers[new_customers$CODE_GENDER != "XNA", ]
new_customers$CODE_GENDER <- factor(new_customers$CODE_GENDER)

# Convert education type to factor with levels across education
new_customers$NAME_EDUCATION_TYPE <- factor(new_customers$NAME_EDUCATION_TYPE
, levels = c(
  "Secondary / secondary special",
  "Higher education",
  "Lower secondary",
  "Incomplete higher",
  "Academic degree"))

# Normalize new customer data using the same scaling as the training data
new_customers_norm <- predict(norm_values, new_customers[, -c(1)])

# Predict risk of new customers
new_customer_predictions <- predict(logistic_model, newdata = new_customers_n
orm, type = "response")

```

```
new_customer_predictions_class <- ifelse(new_customer_predictions > 0.5, 1, 0
)
```

```
# Display predictions for new customers
```

```
new_customer_results <- data.frame(new_customers, Predicted_Risk = new_custom
er_predictions_class)
```

```
head(new_customer_results)
```

```
## ...1 SK_ID_CURR NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REA
LTY
```

```
## 1 1 402254 Cash loans M N
```

```
Y
```

```
## 2 2 440463 Cash loans F N
```

```
Y
```

```
## 3 3 242185 Cash loans F N
```

```
N
```

```
## 4 4 235118 Cash loans M Y
```

```
Y
```

```
## 5 5 407346 Cash loans F N
```

```
Y
```

```
## CNT_CHILDREN AMT_INCOME_TOTAL AMT_CREDIT AMT_ANNUITY AMT_GOODS_PRICE
```

```
## 1 0 247500 746280 59094.0 675000
```

```
## 2 0 112500 589500 32107.5 589500
```

```
## 3 1 112500 272520 16803.0 225000
```

```
## 4 0 157500 533313 37246.5 472500
```

```
## 5 2 112500 283500 22527.0 283500
```

```
## NAME_TYPE_SUITE NAME_INCOME_TYPE NAME_EDUCATION_TYPE
```

```
## 1 Unaccompanied Working Secondary / secondary special
```

```
## 2 Unaccompanied Commercial associate Incomplete higher
```

```
## 3 Unaccompanied Working Secondary / secondary special
```

```
## 4 Unaccompanied Working Higher education
```

```
## 5 Unaccompanied Working Secondary / secondary special
```

```
## NAME_FAMILY_STATUS NAME_HOUSING_TYPE DAYS_BIRTH DAYS_EMPLOYED
```

```
## 1 Separated House / apartment -9889 -2077
```

```
## 2 Married House / apartment -9843 -2772
```

```
## 3 Married House / apartment -10208 -853
```

```
## 4 Married House / apartment -21121 -3561
```

```
## 5 Civil marriage House / apartment -19354 -5103
```

```
## DAYS_REGISTRATION DAYS_ID_PUBLISH OWN_CAR_AGE FLAG_MOBIL FLAG_EMP_PHONE
```

```
## 1 -417 -1342 NA 1 1
```

```
## 2 -524 -2523 NA 1 1
```

```
## 3 -1893 -1946 NA 1 1
```

```
## 4 -7328 -3506 3 1 1
```

```
## 5 -9478 -2873 NA 1 1
```

```
## FLAG_WORK_PHONE FLAG_CONT_MOBILE FLAG_PHONE FLAG_EMAIL OCCUPATION_TYPE
```

```
## 1 0 1 0 0 Laborers
```

```
## 2 0 1 0 0 Cooking staff
```

```
## 3 1 1 0 0 Sales staff
```

```
## 4 0 1 0 1 Drivers
```

```
## 5 0 1 0 0 Laborers
```

##	CNT_FAM_MEMBERS	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY
## 1	1	3	3
## 2	2	2	2
## 3	3	2	2
## 4	2	2	2
## 5	4	2	2
##	WEEKDAY_APPR_PROCESS_START	HOUR_APPR_PROCESS_START	REG_REGION_NOT_LIVE_REGION
## 1	TUESDAY	9	
0			
## 2	THURSDAY	12	
0			
## 3	THURSDAY	13	
0			
## 4	TUESDAY	17	
0			
## 5	FRIDAY	13	
0			
##	REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	REG_CITY_NOT_LIVE_CITY
## 1	0	0	
0			
## 2	0	0	
1			
## 3	0	0	
0			
## 4	0	0	
0			
## 5	0	0	
1			
##	REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	ORGANIZATION_TYPE
## 1	0	0	Other
## 2	1	0	Business Entity Type 2
## 3	1	1	Trade: type 3
## 4	0	0	Business Entity Type 3
## 5	1	0	Business Entity Type 3
##	DAYS_LAST_PHONE_CHANGE	FLAG_DOCUMENT_2	FLAG_DOCUMENT_3
## 1	-785	0	1
## 2	-202	0	0
## 3	-1474	0	1
## 4	-618	0	1
## 5	-510	0	1
##	FLAG_DOCUMENT_5	FLAG_DOCUMENT_6	FLAG_DOCUMENT_7
## 1	0	0	0
## 2	0	0	0
## 3	0	0	0
## 4	0	0	0
## 5	0	0	0
##	FLAG_DOCUMENT_9	FLAG_DOCUMENT_10	FLAG_DOCUMENT_11
## 1	0	0	0

```

## 2          0          0          0          0
## 3          0          0          0          0
## 4          0          0          0          0
## 5          0          0          0          0
##  FLAG_DOCUMENT_13 FLAG_DOCUMENT_14 FLAG_DOCUMENT_15 FLAG_DOCUMENT_16
## 1          0          0          0          0
## 2          0          0          0          0
## 3          0          0          0          0
## 4          0          0          0          0
## 5          0          0          0          0
##  FLAG_DOCUMENT_17 FLAG_DOCUMENT_18 FLAG_DOCUMENT_19 FLAG_DOCUMENT_20
## 1          0          0          0          0
## 2          0          0          0          0
## 3          0          0          0          0
## 4          0          0          0          0
## 5          0          0          0          0
##  FLAG_DOCUMENT_21 AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY
## 1          0          0          0
## 2          0          0          0
## 3          0          NA          NA
## 4          0          0          0
## 5          0          0          0
##  AMT_REQ_CREDIT_BUREAU_WEEK AMT_REQ_CREDIT_BUREAU_MON
## 1          0          0
## 2          0          0
## 3          NA          NA
## 4          0          0
## 5          0          0
##  AMT_REQ_CREDIT_BUREAU_QRT AMT_REQ_CREDIT_BUREAU_YEAR Income_Credit_Ratio
## 1          1          3          0.3316450
## 2          0          2          0.1908397
## 3          NA          NA          0.4128137
## 4          0          0          0.2953238
## 5          0          3          0.3968254
##  Annuity_Income_Ratio Credit_As_Percentage Percent_Days_Employed
## 1          0.2387636          3.015273          0.21003135
## 2          0.2854000          5.240000          0.28162146
## 3          0.1493600          2.422400          0.08356191
## 4          0.2364857          3.386114          0.16859997
## 5          0.2002400          2.520000          0.26366643
##  Income_Per_Person Predicted_Risk
## 1          247500          1
## 2          56250          1
## 3          37500          1
## 4          78750          0
## 5          28125          0

```

```

new_customer_results <- data.frame(
  Customer_ID = new_customers$SK_ID_CURR, # Replace with the actual identifier column if different

```

```

    Prediction = new_customer_predictions_class,
    Probability = new_customer_predictions
)

# Format and display top results for clarity
head(new_customer_results[order(-new_customer_results$Probability), ]) # Top
predictions with high probability

##   Customer_ID Prediction Probability
## 1      402254          1    0.7369819
## 3      242185          1    0.5901293
## 2      440463          1    0.5713890
## 5      407346          0    0.4692277
## 4      235118          0    0.4174471

```