Logistic Regression

Zander Bonnet

2024-05-22

```
app_rec <- read.csv('/Users/zanderbonnet/Desktop/GCU/DSC_520/Data/CreditCard/application
_record.csv')
head(app_rec)
```

```
##        ID CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY CNT_CHILDREN
## 1 5008804           M            Y               Y            0
## 2 5008805           M            Y               Y            0
## 3 5008806           M            Y               Y            0
## 4 5008808           F            N               Y            0
## 5 5008809           F            N               Y            0
## 6 5008810           F            N               Y            0
##   AMT_INCOME_TOTAL     NAME_INCOME_TYPE           NAME_EDUCATION_TYPE
## 1           427500              Working              Higher education
## 2           427500              Working              Higher education
## 3           112500              Working Secondary / secondary special
## 4           270000 Commercial associate Secondary / secondary special
## 5           270000 Commercial associate Secondary / secondary special
## 6           270000 Commercial associate Secondary / secondary special
##     NAME_FAMILY_STATUS NAME_HOUSING_TYPE DAYS_BIRTH DAYS_EMPLOYED FLAG_MOBIL
## 1       Civil marriage  Rented apartment     -12005         -4542          1
## 2       Civil marriage  Rented apartment     -12005         -4542          1
## 3              Married House / apartment     -21474         -1134          1
## 4 Single / not married House / apartment     -19110         -3051          1
## 5 Single / not married House / apartment     -19110         -3051          1
## 6 Single / not married House / apartment     -19110         -3051          1
##   FLAG_WORK_PHONE FLAG_PHONE FLAG_EMAIL OCCUPATION_TYPE CNT_FAM_MEMBERS
## 1               1          0          0                               2
## 2               1          0          0                               2
## 3               0          0          0  Security staff               2
## 4               0          1          1     Sales staff               1
## 5               0          1          1     Sales staff               1
## 6               0          1          1     Sales staff               1
```

```
summary(app_rec)
```

```
##        ID          CODE_GENDER        FLAG_OWN_CAR       FLAG_OWN_REALTY
##  Min.   :5008804   Length:438557      Length:438557      Length:438557
##  1st Qu.:5609375   Class :character   Class :character   Class :character
##  Median :6047745   Mode  :character   Mode  :character   Mode  :character
##  Mean   :6022176
##  3rd Qu.:6456971
##  Max.   :7999952
##   CNT_CHILDREN     AMT_INCOME_TOTAL  NAME_INCOME_TYPE   NAME_EDUCATION_TYPE
##  Min.   : 0.0000   Min.   :  26100   Length:438557      Length:438557
##  1st Qu.: 0.0000   1st Qu.: 121500   Class :character   Class :character
##  Median : 0.0000   Median : 160780   Mode  :character   Mode  :character
##  Mean   : 0.4274   Mean   : 187524
##  3rd Qu.: 1.0000   3rd Qu.: 225000
##  Max.   :19.0000   Max.   :6750000
##  NAME_FAMILY_STATUS NAME_HOUSING_TYPE    DAYS_BIRTH     DAYS_EMPLOYED
##  Length:438557      Length:438557      Min.   :-25201   Min.   :-17531
##  Class :character   Class :character   1st Qu.:-19483   1st Qu.: -3103
##  Mode  :character   Mode  :character   Median :-15630   Median : -1467
##                                        Mean   :-15998   Mean   : 60564
##                                        3rd Qu.:-12514   3rd Qu.:  -371
##                                        Max.   : -7489   Max.   :365243
##    FLAG_MOBIL FLAG_WORK_PHONE    FLAG_PHONE       FLAG_EMAIL
##  Min.   :1    Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1    1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1    Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :1    Mean   :0.2061   Mean   :0.2878   Mean   :0.1082
##  3rd Qu.:1    3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1    Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  OCCUPATION_TYPE    CNT_FAM_MEMBERS
##  Length:438557      Min.   : 1.000
##  Class :character   1st Qu.: 2.000
##  Mode  :character   Median : 2.000
##                     Mean   : 2.194
##                     3rd Qu.: 3.000
##                     Max.   :20.000
```

To make the approved variable I took a handful of variables and used them determine if the application should be approved or not. These variables include the total income, number of children, number of family members, and length of employment.

```
approve <- function(x){
  aprove = 0
  if(x[6] > 190000){
    aprove = aprove + .8
  }
  if(x[12] < 1000){
    aprove = aprove + .3
  }
  aprove = aprove - .1*as.numeric(x[5])
  aprove = aprove + .1*as.numeric(x[18])
  if(x[3] == 'Y'){
    aprove = aprove + .3
  }
  if(aprove >.5){
    return(1)
  }
  else{
    return(0)
  }
}

app_rec$APPROVED = apply(app_rec,1, approve)
sum(app_rec$APPROVED)
```

```
## [1] 148327
```

I then clean the data by first removing the blank occupations, as they are just non-responces. This does not represent unemployed.

```
clean <- app_rec[app_rec$OCCUPATION_TYPE != "",]
```

```
#. A function to calculate the z-score of a variable
zscore <- function(x){
  z <- (x - mean(x))/sd(x)
  return(z)
  }
```

I then calculate the zscore of the numerical predictors so we can remove any outliers.

```
num <- c('CNT_CHILDREN','AMT_INCOME_TOTAL',"DAYS_BIRTH","DAYS_EMPLOYED","CNT_FAM_MEMBER
S")
z <- sapply(clean[,num], zscore)
summary(z)
```

```
##   CNT_CHILDREN     AMT_INCOME_TOTAL    DAYS_BIRTH       DAYS_EMPLOYED
##  Min.   :-0.6675   Min.   :-1.4351   Min.   :-2.85706   Min.   :-6.2115
##  1st Qu.:-0.6675   1st Qu.:-0.5118   1st Qu.:-0.75481   1st Qu.:-0.3692
##  Median :-0.6675   Median :-0.1271   Median : 0.06418   Median : 0.2937
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.6388   3rd Qu.: 0.2576   3rd Qu.: 0.82094   3rd Qu.: 0.7095
##  Max.   :24.1529   Max.   :56.0401   Max.   : 2.09846   Max.   : 1.0874
##  CNT_FAM_MEMBERS
##  Min.   :-1.4016
##  1st Qu.:-0.3218
##  Median :-0.3218
##  Mean   : 0.0000
##  3rd Qu.: 0.7581
##  Max.   :19.1158
```

I then removed any outliers of over 3 SD away from the mean from the predictors. We can do this because the dataset is huge, so we will not lose any information.

```
clean <- na.omit(clean[which(z<=3 & z>=-3),])
clean$FLAG_OWN_CAR <- ifelse(clean$FLAG_OWN_CAR == 'Y', 1, 0)
summary(clean)
```

```
##        ID          CODE_GENDER         FLAG_OWN_CAR    FLAG_OWN_REALTY
##  Min.   :5008806   Length:299498      Min.   :0.0000   Length:299498
##  1st Qu.:5617823   Class :character   1st Qu.:0.0000   Class :character
##  Median :6047764   Mode  :character   Median :0.0000   Mode  :character
##  Mean   :6023047                      Mean   :0.4133
##  3rd Qu.:6448963                      3rd Qu.:1.0000
##  Max.   :7999952                      Max.   :1.0000
##   CNT_CHILDREN   AMT_INCOME_TOTAL  NAME_INCOME_TYPE   NAME_EDUCATION_TYPE
##  Min.   :0.000   Min.   :  27000   Length:299498      Length:299498
##  1st Qu.:0.000   1st Qu.: 135000   Class :character   Class :character
##  Median :0.000   Median : 180000   Mode  :character   Mode  :character
##  Mean   :0.468   Mean   : 194980
##  3rd Qu.:1.000   3rd Qu.: 225000
##  Max.   :2.000   Max.   :6750000
##  NAME_FAMILY_STATUS NAME_HOUSING_TYPE    DAYS_BIRTH     DAYS_EMPLOYED
##  Length:299498      Length:299498      Min.   :-24770   Min.   :-17531
##  Class :character   Class :character   1st Qu.:-17488   1st Qu.: -3511
##  Mode  :character   Mode  :character   Median :-14622   Median : -1917
##                                        Mean   :-14826   Mean   : -2626
##                                        3rd Qu.:-11929   3rd Qu.:  -919
##                                        Max.   : -7489   Max.   :   -12
##    FLAG_MOBIL FLAG_WORK_PHONE    FLAG_PHONE       FLAG_EMAIL
##  Min.   :1    Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1    1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1    Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :1    Mean   :0.2482   Mean   :0.2861   Mean   :0.1175
##  3rd Qu.:1    3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1    Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  OCCUPATION_TYPE    CNT_FAM_MEMBERS    APPROVED
##  Length:299498      Min.   :1.000   Min.   :0.0000
##  Class :character   1st Qu.:2.000   1st Qu.:0.0000
##  Mode  :character   Median :2.000   Median :0.0000
##                     Mean   :2.253   Mean   :0.4132
##                     3rd Qu.:3.000   3rd Qu.:1.0000
##                     Max.   :4.000   Max.   :1.0000
```

I will use logistic regression to predict the approval of a loan based on the given factors from the loan application.

I split the data into a training and testing set, where 70% of the data is in the training set and 30% is in the testing set.

```
set.seed(100)
trainind <- sample(seq_len(nrow(clean)), size = nrow(clean)*.7)
train <- clean[trainind,]
test <- clean[-trainind,]
```

Here I make the model using occupation type, gender, total income, number of children, and number of family members.

```
mod <- glm(APPROVED~ OCCUPATION_TYPE +
             CODE_GENDER  + AMT_INCOME_TOTAL + CNT_CHILDREN + CNT_FAM_MEMBERS,
           data = train,family = binomial)
summary(mod)
```

```
##
## Call:
## glm(formula = APPROVED ~ OCCUPATION_TYPE + CODE_GENDER + AMT_INCOME_TOTAL +
##     CNT_CHILDREN + CNT_FAM_MEMBERS, family = binomial, data = train)
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -2.048e+00  3.213e-02 -63.729  < 2e-16 ***
## OCCUPATION_TYPECleaning staff     -1.208e+00  5.032e-02 -24.011  < 2e-16 ***
## OCCUPATION_TYPECooking staff      -6.835e-01  3.809e-02 -17.944  < 2e-16 ***
## OCCUPATION_TYPECore staff         -3.713e-01  2.390e-02 -15.535  < 2e-16 ***
## OCCUPATION_TYPEDrivers             3.115e-01  2.845e-02  10.950  < 2e-16 ***
## OCCUPATION_TYPEHigh skill tech staff -2.554e-01  2.846e-02  -8.974  < 2e-16 ***
## OCCUPATION_TYPEHR staff           -1.704e-01  9.887e-02  -1.723  0.08481 .
## OCCUPATION_TYPEIT staff           -5.867e-01  1.063e-01  -5.522 3.36e-08 ***
## OCCUPATION_TYPELaborers           -5.206e-01  2.328e-02 -22.357  < 2e-16 ***
## OCCUPATION_TYPELow-skill Laborers -9.918e-01  6.102e-02 -16.254  < 2e-16 ***
## OCCUPATION_TYPEManagers            1.991e-02  2.471e-02   0.806  0.42033
## OCCUPATION_TYPEMedicine staff     -4.928e-01  3.133e-02 -15.728  < 2e-16 ***
## OCCUPATION_TYPEPrivate service staff -5.766e-02  4.753e-02  -1.213  0.22508
## OCCUPATION_TYPERealty agents      -2.660e-01  8.221e-02  -3.236  0.00121 **
## OCCUPATION_TYPESales staff        -4.213e-01  2.420e-02 -17.413  < 2e-16 ***
## OCCUPATION_TYPESecretaries        -3.171e-01  6.222e-02  -5.096 3.47e-07 ***
## OCCUPATION_TYPESecurity staff     -5.368e-01  3.616e-02 -14.843  < 2e-16 ***
## OCCUPATION_TYPEWaiters/barmen staff -9.270e-01  7.917e-02 -11.708  < 2e-16 ***
## CODE_GENDERM                       1.243e+00  1.185e-02 104.972  < 2e-16 ***
## AMT_INCOME_TOTAL                   2.703e-06  5.333e-08  50.685  < 2e-16 ***
## CNT_CHILDREN                      -3.245e-01  1.509e-02 -21.502  < 2e-16 ***
## CNT_FAM_MEMBERS                    5.066e-01  1.230e-02  41.192  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 284238  on 209647  degrees of freedom
## Residual deviance: 250174  on 209626  degrees of freedom
## AIC: 250218
##
## Number of Fisher Scoring iterations: 4
```

The model results in all the factors being very significant with extremely low p-values.

```
anova(mod, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: APPROVED
##
## Terms added sequentially (first to last)
##
##
##                   Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                              209647     284238
## OCCUPATION_TYPE  17  14354.8    209630     269883 < 2.2e-16 ***
## CODE_GENDER       1  14366.1    209629     255517 < 2.2e-16 ***
## AMT_INCOME_TOTAL  1   2514.6    209628     253003 < 2.2e-16 ***
## CNT_CHILDREN      1   1081.5    209627     251921 < 2.2e-16 ***
## CNT_FAM_MEMBERS   1   1747.6    209626     250174 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To check the significance of all the predictors I use ANOVA to show that all the predictors are significant in predicting the response.

Looking at the training data we can see that the model is 69% accurate in predicting approval in the training data. By looking at the ROC curve we can see that the model does not perform extremely well with only 72.9% of the area being under the curve. We can also see to get our true positive rate to about 80% our false positive rate would have to be about 60%, so the model does not perform great.

```
library(ROCR)
library(Metrics)


preds <- predict(mod, type = 'response')
glm.pred <- rep("0", nrow(train))
glm.pred[preds > .5] = "1"

table(glm.pred, train$APPROVED)
```

```
##
## glm.pred     0     1
##        0 98802 40393
##        1 24284 46169
```

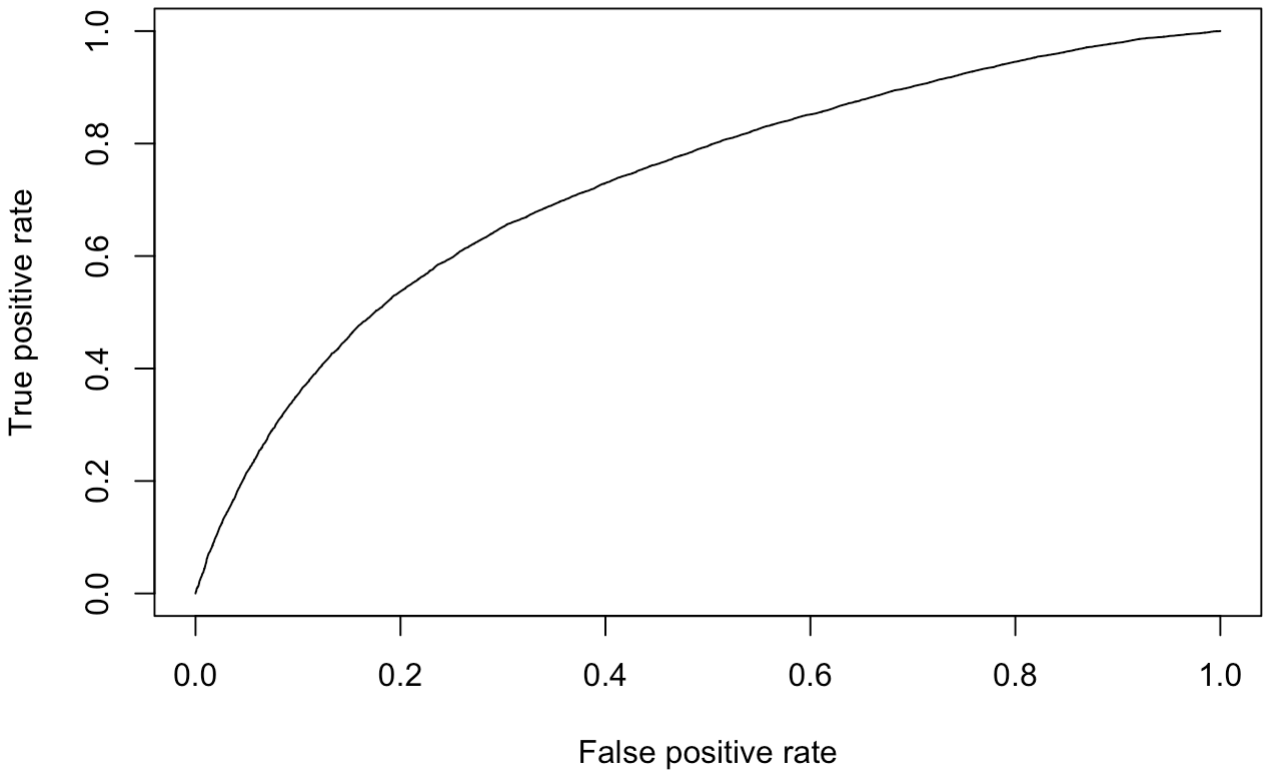
```
mean(glm.pred == train$APPROVED)
```

```
## [1] 0.6914972
```

```
pr <- prediction(preds ,train$APPROVED)
perf <- performance(pr,measure = "tpr",x.measure = "fpr")
auc(train$APPROVED,preds)
```

```
## [1] 0.7288214
```

```
plot(perf)
```




```
library(regclass)
```

```
## Loading required package: bestglm
```

```
## Loading required package: leaps
```

```
## Loading required package: VGAM
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

```
## Loading required package: rpart
```

```
## Loading required package: randomForest
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
## Important regclass change from 1.3:
## All functions that had a . in the name now have an _
## all.correlations -> all_correlations, cor.demo -> cor_demo, etc.
```

```
VIF(mod)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## OCCUPATION_TYPE  1.573815 17        1.013428
## CODE_GENDER      1.445431  1        1.202261
## AMT_INCOME_TOTAL 1.116029  1        1.056423
## CNT_CHILDREN     4.699554  1        2.167845
## CNT_FAM_MEMBERS  4.709564  1        2.170153
```

When looking at the VIF we see that there is no multicoliniarity in the predictors, There is a possibility between the number of children and family members, but the VIF is not large enough to make us definitively say that that is the case. Since there is no strong evidence of multicolliniarity the model can remain as is.

```
pred <- predict(mod, newdata = test, type = 'response')
pred <- ifelse(pred >.5, 1,0)

table(pred, test$APPROVED)
```

```
##
## pred     0     1
##    0 42228 17267
##    1 10428 19927
```

```
mean(pred == test$APPROVED)
```

```
## [1] 0.6917641
```

When using the model to predict the testing data we get a 68.9% success rate. This shows that the model is not over fit to the training data, because the model performed very similarly on on the two independent sets. The model seems to have a higher rate of false negatives then false possitives. This means that the model will not award loans to undeserving applicants as often as not giving loans to qualified applicants.

```
pred <- predict(mod, newdata = clean, type = 'response')
pred <- ifelse(pred >.5, 1,0)

table(pred, clean$APPROVED)
```

```
##
## pred      0      1
##    0 141030  57660
##    1  34712  66096
```

```
mean(pred == clean$APPROVED)
```

```
## [1] 0.6915772
```

When using the model to predict the entire cleaned data set we get again a very similar result of about a 69% success rate. We can also see that the model has a high rate of false negatives, shown in the top right of the table.

In the end the model does not perform very well to predict the outcome of the loan application with a success rate of about 70%. It does lean towards resulting false negatives though so it is less likely to give loans to unqualified applicants. This does mean that the model will not award loans to qualified applicants more often though. In a risk assessment for the business this may be the more desirable approach.

To improve this model I might consider diving deeper into the different predictive factors in the model. I might use backwards selection, starting with all the predictors, to find the most effective predictors. I might also chose to use transformations on some of the predictors, like income, to see how that might effect the efficacy. Especially because some of th emore extreme values of that variable were removed by the outlier treatment. By doing this we might be able to get more accurate based on the income predictor.

Reference

Credit Card Approval Prediction. (2020). Kaggle [Dataset]. https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction (https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction).