# Logistic Regression to Predict Likelihood of Loan Repayment

Andrew vanderWilden

October 29, 2019

## 1. Abstract

This report uses data on customer financial and demographic information. In the report, a Logistic Regression model is used to predict the credit worthiness of a customer. We found `AGE`, `DISP` (Disposable Income), `PHON` (Presence of a Phone in the House), and `AES` (Applicant Employment Status) to be significant predictors of credit worthiness. Additionally we found using a threshold of 18% to decide if a customer is a bad credit risk returned the highest net profits for the bank.

## 2. Introduction

**Orientation Material** Loan companies offer loans to customers in need of money with the expectation the customer will be able to pay back the balance of the loan plus interest over a specified term length. A customer that defaults on a loan can be extremely expensive as profits from customers who successfully pay the balance of their loan only account for fractions of the total amount of an unrecoverable loan. Put another way, it is paramount to determine the credit worthiness of potential customers before offering a loan.

This report attempts to evaluate credit worthiness of customers using data from the book *Credit Scoring and its Applications* by Lyn C. Thomas, David B. Edelman, and Jonathan N. Crook. The data is estimated to have been collected in the year 2000.

The data includes information about customers' demographic traits, income, and outgoing expenses.

**Key Aspects** This report attempts to fit a model using Logistic Regression to predict if a customer will fail to repay a loan. The model returns the likelihood of a customer defaulting on a loan which is then used to classify the customer as either a good or bad credit risk.

**Plan for the Rest of the Report** The outline for the remainder of the report is as follows. In section 3, we present the most important characteristics of the data and the relationships between predictor variables and credit worthiness. In section 4, the model selection process and the following interpretation will be discussed. Concluding remarks can be found in section 5 with details to follow in the Appendix.

## 3. Data Characteristics

The data are cross-sectional and describe credit information for 900 customers at a non-specified time. The data includes information on the following 29 variables:

| Item | Variable | Definition |
|---|---|---|
| 1 | DOB | Year of birth |
| 2 | NKID | Number of Children |
| 3 | DEP | Number of Other Dependents |
| 4 | PHON | Is there a Home Phone |
| 5 | SINC | Spouse's Income |
| 6 | AES | Applicant Employment Status |
| 7 | DAINC | Applicant's Income |
| 8 | RES | Residential Status |
| 9 | DHVAL | Value of Home |
| 10 | DMORT | Mortgage Balance Outstanding |
| 11 | DOUTM | Outgoing Payments on Mortgage or Rent |
| 12 | DOUTL | Outgoing Payments on Loans |
| 13 | DOUTHP | Outgoing Payments on Hire Purchase |
| 14 | DOUTCC | Outgoing Payments on Credit Cards |
| 15 | BAD | Good/Bad Credit Indicator |

See Appendix Section A for the coding of values for the categorical variable `AES`.

The variable of interest is the variable `BAD`. A value of 1 indicates the customer did not repay the loan while a value of 0 indicates the customer successfully repaid the loan. Of the 900 customers in the data set, 240, or 26.67% failed to pay the balance of their loan.

The variable `DOB` was used to create a new variable `AGE`. `AGE` is calculated by subtracting 1900 + `DOB` from the year 2000.

The below table shows summary statistics for all numerical variables in the data set.

| | Mean | Median | Standard Deviation | Minimum | Maximum |
|---|---|---|---|---|---|
| AGE | 49.58 | 45 | 15.68 | 0 | 97 |
| NKID | 0.61 | 0 | 1.00 | 0 | 5 |
| DEP | 0.04 | 0 | 0.22 | 0 | 2 |
| SINC | 1956.01 | 0 | 4693.01 | 0 | 50000 |
| DAINC | 20925.90 | 18928 | 15949.20 | 0 | 64800 |
| DHVAL | 15548.97 | 0 | 21007.56 | 0 | 64928 |
| DMORT | 10263.39 | 0 | 18054.33 | 0 | 64000 |
| DOUTM | 326.58 | 240 | 423.91 | 0 | 3800 |
| DOUTL | 104.13 | 0 | 283.16 | 0 | 5200 |
| DOUTHP | 27.48 | 0 | 118.15 | 0 | 1600 |
| DOUTCC | 39.22 | 0 | 180.13 | 0 | 2800 |

Because the variable `DOB` was coded using the value 99 for year of birth unknown, the variable `AGE` is smaller than the true mean because there are 4 observations that appear to have an age of 1.

It is also worth noting that the documentation accompanying the data set did not indicate what the time scales of the income variables or outgoing payment variables are. Based on the summary statistics it appears as though all outgoing payments appear to be on a monthly scale whereas the income variables appear to be on a yearly scale. For the purposes of the rest of this report, they will be treated as such.

Individually the variable for outgoing monthly payments may not be useful for prediction. With the exception of the variable `DOUTM`, more than half of all customers have no payments in each individual category. It likely makes more sense to combine all of these payments into one new variable called `OUTPAY`.

We can see the summary statistics for the variable `OUTPAY` below:

```
   Mean   Median Std. Dev      Min      Max
  497.4    380.0    615.6      0.0   5400.0
```

Even after this transformation, there are still 296 customers that apparently have no monthly expenses at all which suggests an error in data collection or at the least it is extremely anomalous.

Additionally, it would make sense to consider the overall income level of an entire household, rather than individually. We created a new variable called `HHINC` by combining `SINC` and `DAINC`. We can see the summary statistics for the variable `HHINC` below:

```
    Mean   Median Std. Dev      Min      Max
22881.91 21000.00 16649.03     0.00 88000.00
```

Again, it is noteworthy that 136 customers appear to have no household income. Of this group, 41 appear to be retired and may be relying on the value of owned assets to be able to repay a loan. 48 potential customers are students with no income. We can see from the below summary information that as one would expect, people who have no household income are likely to have difficulty repaying a loan:

```
Total  Good   Bad
  136    66    70
```

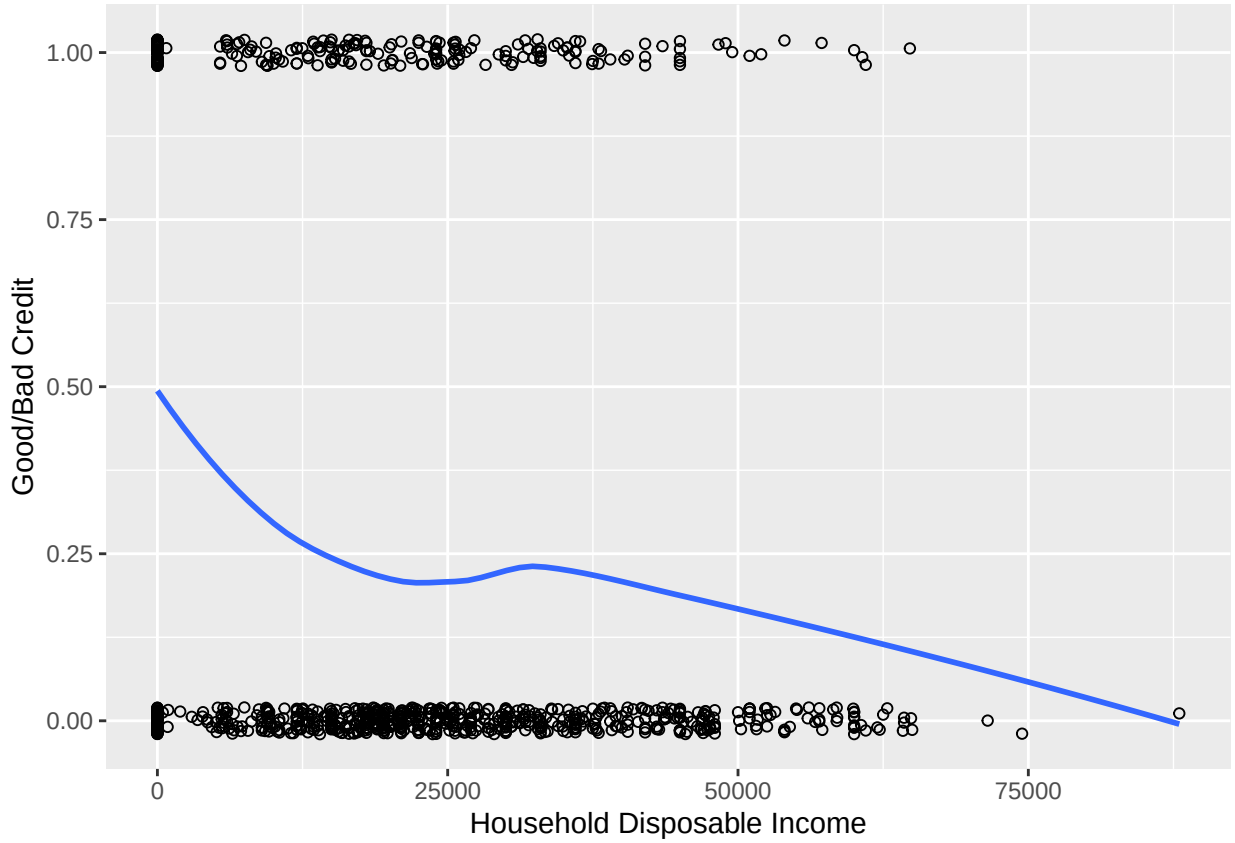
Strangely, 23 people appear to be working jobs in the military, private sector, public sector, or are self-employed but seemingly are doing so for free. This suggests an error in data collection or coding.

To gain a better understanding of a customer's ability to repay a loan, calculating a household's disposable income gives an estimate of available funds a customer will have after accounting for all outgoing expenses. We created the variable `DISP` to represent this value. It was calculated by multiplying the total monthly outgoing payments by 12 to account for the monthly time scale, and subtracting that value from the total household income. We can see the summary statistics below:

```
     Mean    Median  Std. Dev       Min       Max
 16913.05  15906.00  15011.22 -52936.00  65000.00
```

The below plot shows the relationship between household income and credit status.

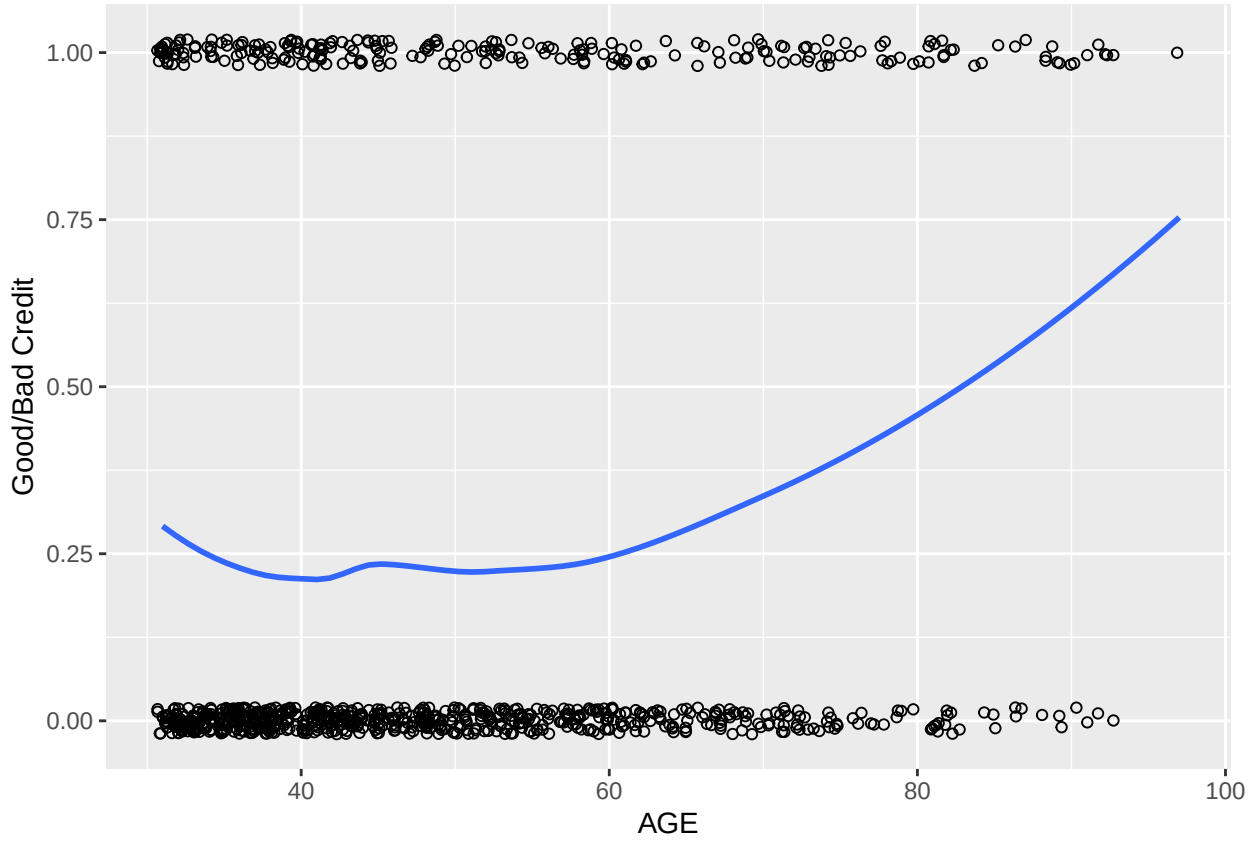
```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




The blue line shows the general relationship between the two variables. As Household Disposable Income increases, the less likely the customer is to be a bad credit risk. This indicates the variable `DISP` is likely an important predictor of credit worthiness.

The variable `AGE` appears to be a significant predictor of credit worthiness as well. As we can see from the trend line on the plot below, older people are associated with greater credit risk. This is likely because they do not have a reliable stream of income.

```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




The variable PHON appears to be a significant predictor as well:

```
No Phone.0    Phone.1
 0.3478261  0.2574257
```

Having a phone in the house appears to be significantly correlated with credit worthiness.

The variable AES is likely also a useful predictor of credit worthiness. The below table shows the breakdown of good and bad credit by employment status as well as the total percentage for each group:

```
              B          E          M N           P          R          T
BAD   8.0000000 28.0000000  2.0000000 0  90.0000000 41.0000000 23.0000000
GOOD 15.0000000 61.0000000 12.0000000 3 304.0000000 45.0000000 62.0000000
%BAD  0.3478261  0.3146067  0.1428571 0   0.2284264  0.4767442  0.2705882
             U           V          W   Z
BAD  4.0000000  28.0000000 13.0000000 3.0
GOOD 3.0000000 136.0000000 16.0000000 3.0
%BAD 0.5714286   0.1707317  0.4482759 0.5
```

We can observe there are clear differences in credit worthiness among the categories with a significant number of observations. For purposes of model building, the categories Z and N, which correspond to no response or other were pooled together. Consideration was given to also adding the category U however we felt knowing a potential customer is unemployed is an important piece of information

# 4. Model Selection and Interpretation

Based on the above Data Characteristics section, it has been established there are clear correlations and patterns between the credit indicator, and many of the predictor variables.

In this section we summarize these relationships using regression modeling. We also explain the ways in which we manipulated the data during our selection process.

Based on our investigation of the data, we recommend a Logistic regression model using a Logit link function to estimate credit worthiness. The variables used to create the regression model are: `AGE`, `DISP` (Disposable Income), `PHON` (Presence of a Phone in the House), and `AES` (Applicant Employment Status). The variable `AES` was transformed into a derived grouped variable. The variable `DISP` was scaled so an increase of 1 indicates an increase of $1000 of disposable income.

When using the model to make classifications of new customers, we recommend using a probability threshold of 18% to classify someone as a bad credit risk.

Additionally two indicator variables were used to account for anomalies in the data described in the above section. The variable `AGE_UNKN` has a value of 1 for any observation which reported a `DOB` of 99 and a 0 for all other observations. The variable `HHINC_UNKN` has a value of 1 for any observation that had no household income and a value of 0 for all other observations.

The model was built using all 900 observations from the data set and tested using K-fold cross validation with K equal to 10.

The model was fit using an iteratively weighted least squares algorithm and the following table shows the value of the estimated coefficients and their standard errors.

| | Estimate | Std. Error |
|---|---|---|
| (Intercept) | -0.783 | 0.606 |
| AGE | 0.018 | 0.007 |
| AGE_UNKN | 1.673 | 1.236 |
| DISP.scl | -0.016 | 0.007 |
| HHINC_UNKN | 0.990 | 0.273 |
| emp.statE | -0.301 | 0.504 |
| emp.statM | -1.508 | 0.907 |
| emp.statOther | -1.257 | 0.957 |
| emp.statP | -0.618 | 0.459 |
| emp.statR | -0.671 | 0.568 |
| emp.statT | -1.064 | 0.551 |
| emp.statU | -0.165 | 0.929 |
| emp.statV | -0.950 | 0.490 |
| emp.statW | -0.183 | 0.600 |
| PHON | -0.439 | 0.257 |

The category `B`(public sector) is taken to be the base level for the regression.

The coefficient of the variable `DISP.scl` is equal to -0.016. We can interpret this to mean an increase of $1,000 in disposable income decreases the odds of being classified as a bad credit risk by a multiplicative factor of $\exp(-0.016) = 0.984$.

As a simple example, if the previous odds were equal to 0.5, new odds $= 0.5 \times 0.984 = 0.492$

The previous probability was:

$$\pi = \frac{0.5}{1.5} = 0.3333$$

The new probability would be:

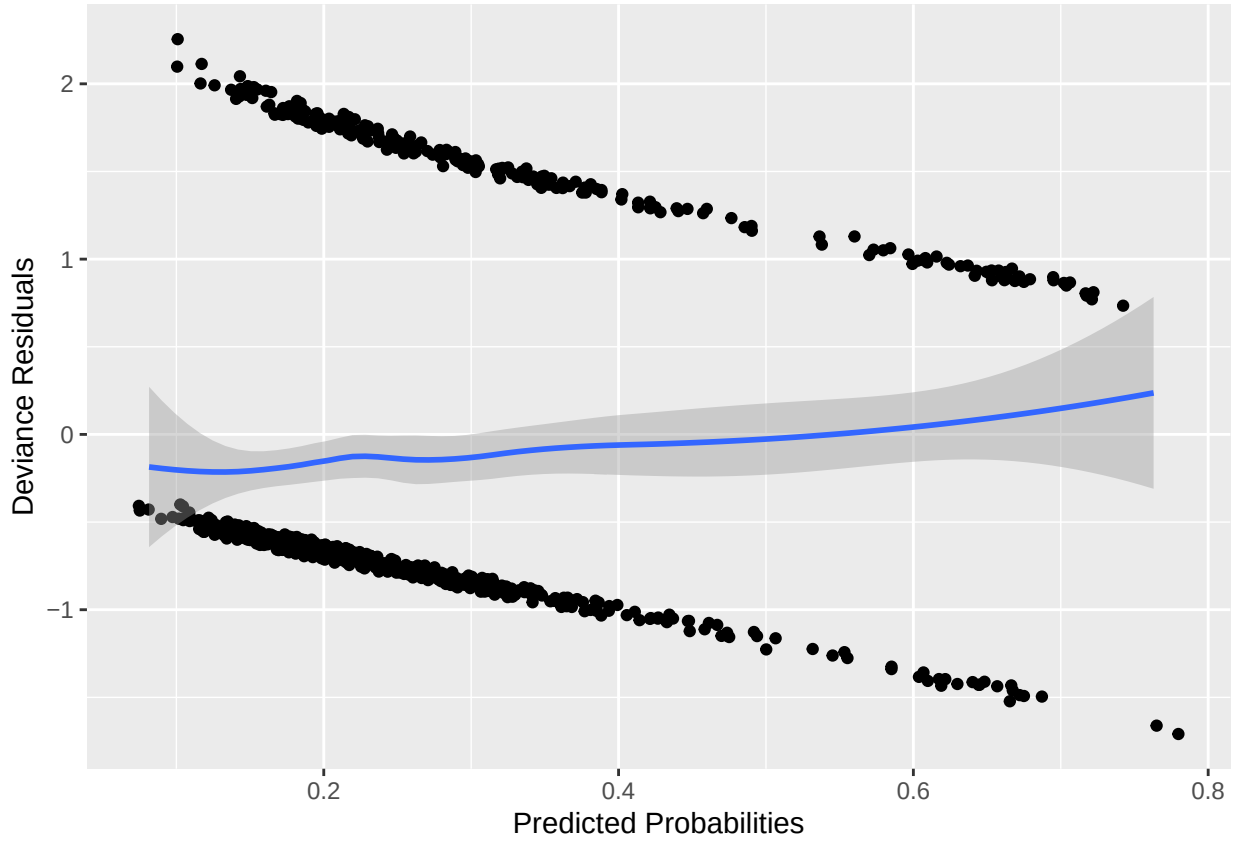
$$\pi = \frac{0.492}{1.492} = 0.3297$$

This is an decrease of roughly 0.36%.

## Discussion of Model and Selection Criteria for Goodness of Fit

The residuals for our recommended model did not show significant patterns. The following graph shows the deviance residuals against the predicted probabilities.

```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




This is a typical plot for logistic regression models. The blue line shows the overall estimate of the pattern of residuals as the predicted probability increases. The line is reasonably close to flat indicating no issues with the residuals. Towards the two ends of the graph the grey area widens to indicate the increasing uncertainty however this is to be expected as we have fewer observations with which to estimate the line.

Typical in classification models, the criterion used to select the optimal model is a combination of accuracy, sensitivity, specificity, and precision of classifiers. The below statistics show the summary values using a classifying threshold of 42% which maximized out accuracy.

| | Accuracy | Precision | Sensitivity | Specificity |
|---|---|---|---|---|
| Whole Sample | 0.7511 | 0.5741 | 0.2583 | 0.9303 |
| Cross-Validated | 0.7344 | 0.5172 | 0.2354 | 0.9193 |
| Difference | 0.0167 | 0.0568 | 0.0229 | 0.0110 |

We felt this was not the correct way to measure accuracy in this application and therefore chose to prioritize potential profit for the bank which will be detailed below.

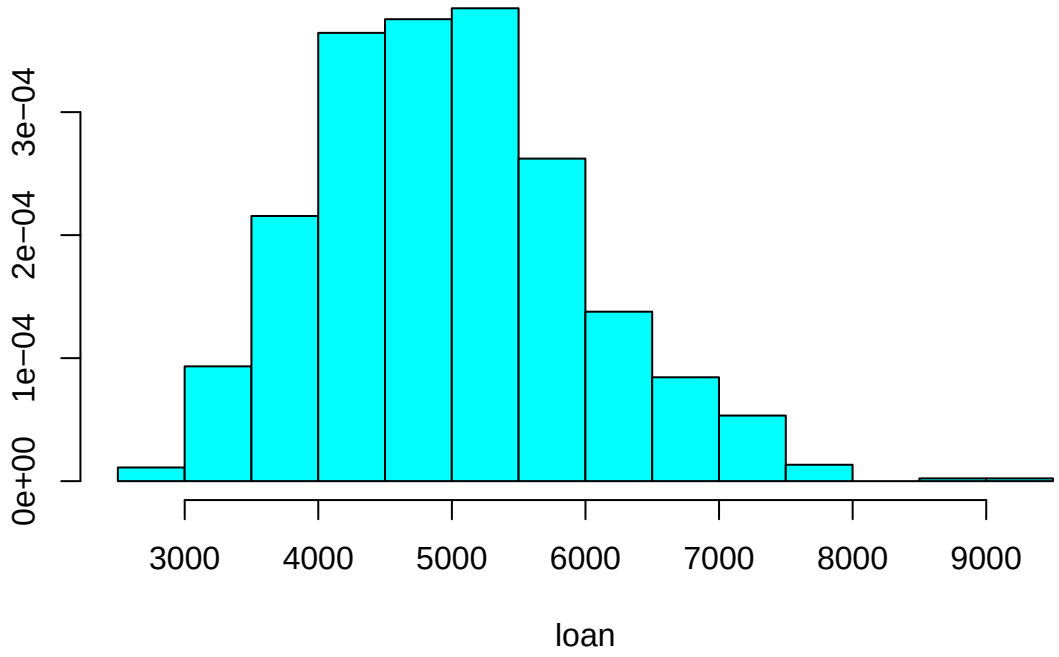
Another measure of accuracy used in logistic regression is the Hosmer Lemeshow statistic. The smaller the statistic is, the better the goodness of fit. The HL stat for our model and its associated p-value are shown below:

```
 HL Stat.   P-Value
4.4529540 0.8141213
```

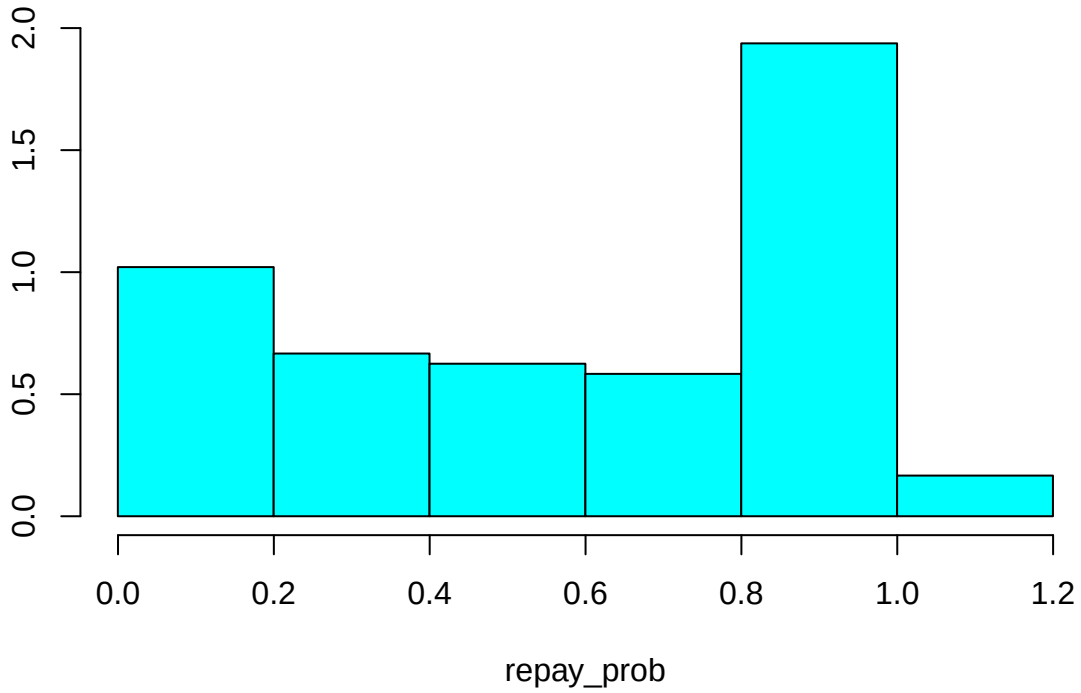
In the model-building process we were able to build models with lower HL statistics but they did not perform well when looking at classification rates or profit maximization.

The primary criteria used to choose our model was a simulation of potential profits or losses a bank could expect using our classifier.

Because the data did not come with any loan amounts we decided to randomly assign loan values using a log-normal distribution with a mean of $5,000 and a standard deviation of 1000. The below chart shows the distribution of loans to give a sense of the relative risk profile.




Banks do not give out loans without collateral. To assume every misclassified bad credit risk would not pay back any of the money is not a realistic situation. To model the percentage of a loan that would be recoverable we decided to randomly assign values from a U-shaped beta distribution with a mean of 0.6. The U-Shape tries to account for the few people who truly have no ability to repay the loan as well as partial collections by means of collateral or partial repayment. The distribution of probabilities can be seen below:




The simulation was performed assuming the bank makes a 10% profit on every loan that is repayed in full. In a real life example, loan amounts and repay probabilities would be tied to the very model we are building and therefore would perform at an even higher level. The results of the simulation can be seen below.The profits reported are an average of 10 runs of the simulation, reassigning the recoverable amount of a loan to different loan amounts. The full simulation can be seen in Appendix section B.

```
   Threshold     Mean
1       0.15 14816.15
2       0.16 20934.38
3       0.17 27978.68
4       0.18 35690.91
5       0.19 30433.24
6       0.20 31154.97
7       0.21 23239.34
8       0.22 26850.56
9       0.23 27597.73
10      0.24 28448.23
11      0.25 23955.56
```

Using a threshold value of 18% consistently resulted in the highest net profits for the bank. For our model, this resulted in a net profit of $35,690.91. If we were to give out loans to every customer and were still able to make partial recoveries, the net loss for each threshold would be:

```
   Threshold      Mean
1       0.15 -139997.6
2       0.16 -139441.4
```

```
3       0.17 -138801.0
4       0.18 -138099.8
5       0.19 -138577.8
6       0.20 -138512.2
7       0.21 -139231.8
8       0.22 -138903.5
9       0.23 -138835.6
10      0.24 -138758.3
11      0.25 -139166.7
```

Using the model resulted in a net difference of $173,790 for just these 900 customers.

The model we used was able to outperform all other models we tested using the same simulation technique.

## 5. Summary and Concluding Remarks

In trying to predict credit worthiness of potential customers, we found the variables `AGE`, `DISP` (Disposable Income), `PHON` (Presence of a Phone in the House), and `AES` (Applicant Employment Status) to be significant predictors in our logistic regression model. We found using a threshold of 18% to classify customers as good or bad credit risks resulted in the highest net profits. It is likely worth exploring if a different type of classification model would outperform a logistic regression model using techniques such as K-nearest-neighbors or Classification Trees.

## Appendix

### Section A

| Item | Category | Definition |
|---|---|---|
| 1 | V | Government |
| 2 | W | Housewife |
| 3 | M | Military |
| 4 | P | Private Sector |
| 5 | B | Public Sector |
| 6 | R | Retired |
| 7 | E | Self-Employed |
| 8 | T | Student |
| 9 | U | Unemployed |
| 10 | N | Others |
| 11 | Z | No Response |

### Section B

```
thresholds <- seq(0.15, 0.25, by = 0.01)
p <- predict(fmod, type = "response")

df$loan = rep(0,length(df$BAD))
loan = rlnorm(n=900, location, shape)
df$loan = sample(loan, replace = FALSE)
```

```
for(i in 1:10){
  y = 1

  for(thr in thresholds){
    set.seed(17*i)

    df$recoup = rep(0,length(df$BAD))

    repay_prob = rbeta(240,.6,.4)
    df$recoup[idx.BAD] = sample(repay_prob, replace = FALSE)

    df$PC = rep(0,length(df$BAD))
    df$prof_loss = rep(0, length(df$BAD))
    df$PC = ifelse(p > thr, 1, 0)

    df$prof_loss = ifelse(df$PC == 1,
                          0, ifelse(df$BAD == 1,
                          (df$loan - (df$loan * df$recoup)) * (-1),
                          df$loan * 0.1))
    result[y,(i+1)] = sum(df$prof_loss)
    y = y+1
  }

}
result$Mean = rowMeans(result[,-1])
profs = result[,c(1,12)]
```

## Score Function

```
score <- function(newdata){
  db <- newdata

  db$OUTPAY = db$DOUTM + db$DOUTL + db$DOUTHP + db$DOUTCC

  db$HHINC = db$SINC + db$DAINC

  db$AGE <- 2000 - (1900 + db$DOB)

  idx.AGE <- which(db$AGE == 1)

  db$AGE[idx.AGE] <- 0

  db$AGE_UNKN <- rep(0, length(db$AGE))
  db$AGE_UNKN[idx.AGE] <- 1

  idx.HHINC <- which(db$HHINC == 0)
  db$HHINC_UNKN <- rep(0, length(db$HHINC))
  db$HHINC_UNKN[idx.HHINC] <- 1

  db$emp.stat <- fct_collapse(db$AES, "Other" = c("N", "Z"))
```

```
  db$emp.stat2<- fct_collapse(db$AES, "Other" =c("M","N","U","Z"))

  db$DISP <- db$HHINC - (12 * db$OUTPAY)
  db$DISP.scl = db$DISP/1000

  p <- predict(fmod, newdata = db, type = "response")
  ans <- ifelse(p > 0.18, 1, 0)
  return(ans)
}
```