STAT 528 midterm exam

Solution Set

Due at 11:59 PM on 3/10/2023

Analyze the SBA loans dataset

This dataset contains 899164 observations and 27 columns. This is historical data about actual business loans covered by the Small Business Administration (SBA) primarily from years 1970-2013 with emphasis on whether those businesses defaulted (charged off) or not (paid in full) on those loans. The observations are small businesses that seek loans to fund their operations, start-up costs, materials, payroll, rent, etc. The SBA works with banks by guaranteeing a portion of the loan to relieve banks of assuming all financial risk.

This will load in the data:

```
library(data.table)
sba<-fread("https://uofi.box.com/shared/static/vi37omgitiaa2yyplrom779qvwk1g14x.csv",
                    header = TRUE, stringsAsFactors = FALSE)
dim(sba)
```

```
## [1] 899164     27
```

Your assignment is to analyze this SBA dataset. Your analysis needs to be interesting and well motivated. Any final model reported needs to be justified, properly validated, and well-fitting. You need to check modeling assumptions for any final models that you report. You are allowed to consider subsets of the data as long as subsetting is well motivated. You are allowed to transform variables and create new variables provided that these manipulations are well motivated. Your analysis should be multifaceted, interesting relationships between variables should be reported. You are allowed to use materials from outside this course provided that you have a good reason for doing so and have considered the materials in this course (for example, if you consider flexible machine learning methods then you need to consider interaction terms in the glms). You are encouraged to add outside variables that may be important (economic measures for example, there have been several economic downturns over the range of data collection).

Point on selection bias: It is believed that the inclusion of loans with disbursement dates after 2010 would provide greater weight to those loans that are charged off versus paid in full. More specifically, loans that are charged off will do so prior to the maturity date of the loan, while loans that will likely be paid in full will do so at the maturity date of loan (which would extend beyond the dataset ending in 2014). Since this dataset has been restricted to loans for which the outcome is known, there is a greater chance that those loans charged off prior to maturity date will be included in the dataset, while those that might be paid in full have been excluded. It is important to keep in mind that any time restriction on the loans included in the data analyses could introduce selection bias, particularly toward the end of time period. This may impact the performance of any predictive models based on these data.

The original source for this data is "Should this loan be approved or denied?: A Large dataset with class assignment guidelines" by Min Li, Amy Mickel, and Stanley Taylor (https://www.tandfonline.com/doi/full/10.1080/10691898.2018.1434342). You are not allowed to copy the analyses in this reference. However, you can compare your analyses to those conducted in this reference.

You should save your midterm as **netid_midterm** and it should be stored in a directory titled **midterm**. Do not include the data set with your submission.

Here is a description of the variables:

Variable name	Data type	Description of variable
LoanNr_ChkDgt	Text	Identifier–Primary key
Name	Text	Borrower name
City	Text	Borrower city
State	Text	Borrower state
Zip	Text	Borrower zip code
Bank	Text	Bank name
BankState	Text	Bank state
NAICS	Text	North American industry classification system code
ApprovalDate	Date/Time	Date SBA commitment issued
ApprovalFY	Text	Fiscal year of commitment
Term	Number	Loan term in months
NoEmp	Number	Number of business employees
NewExist	Text	1 = Existing business, 2 = New business
CreateJob	Number	Number of jobs created
RetainedJob	Number	Number of jobs retained
FranchiseCode	Text	Franchise code, (00000 or 00001) = Nofranchise
UrbanRural	Text	1 = Urban, 2 = rural, 0 = undefined
RevLineCr	Text	Revolving line of credit: Y = Yes, N = No
LowDoc	Text	LowDoc Loan Program: Y = Yes, N = No
ChgOffDate	Date/Time	The date when a loan is declared to be in default
DisbursementDate	Date/Time	Disbursement date
DisbursementGross	Currency	Amount disbursed
BalanceGross	Currency	Gross amount outstanding
MIS_Status	Text	Loan status charged off = DCHGOFF, Paid in full = PIF
ChgOffPrinGr	Currency	Charged-off amount
GrAppv	Currency	Gross amount of loan approved by bank
SBA_Appv	Currency	SBA's guaranteed amount of approvedloan

(You may want to use regular expressions and pattern replacement functions such as `gsub` to convert currency variables to numeric.)

Here is a description of the first two digits of the NAICS classifications:

Sector	Description
11	Agriculture, forestry,fishing and hunting
21	Mining, quarrying, and oil and gas extraction
22	Utilities
23	Construction
31–33	Manufacturing
42	Wholesale trade
44–45	Retail trade
48–49	Transportation and warehousing
51	Information
52	Finance and insurance
53	Real estate and rental and leasing
54	Professional, scientific, and technical services
55	Management of companies and enterprises
56	Administrative and support and waste management and remediation services
61	Educational services
62	Health care and social assistance
71	Arts, entertainment, and recreation
72	Accommodation and food services
81	Other services (except public administration)
92	Public administration

Analysis of the SBA loans dataset

This solution set is one of several possible reasonable analyses. It will be organized as follows:

- We will first clean the data set.
- We will then consider data exploration and investigate interesting variables.
- We will then fit and examine a logistic regression model.
- We will then present results, validation, and comparisons with a simple neural network and random forest.
- Concluding remarks will then be provided.

The analysis presented here will be from the perspective of an SBA official who has basic covariates on businesses, including term length, but does not have any future information on the businesses (jobs created and jobs retained variables are ignored). This hypothetical SBA official will be interested in simple models that are well-fitting and offer satisfactory predictive/classification performance.

We load in the following packages:

```
library(tidyverse)
library(ggplot2)
library(lubridate)
library(heatmapFit)
library(caret)
library(pROC)
library(nnet)
library(randomForest)
```

Data Cleaning

We start by performing some cleaning on the variables that make up the covariates in our model. First, since our response variable is whether the customer defaulted on the loan or not, the variable `MIS_Status` is our response variable. We also count the number of rows that do not have a value for `MIS_Status` as 1997 rows.

```
sba <- sba %>% mutate(MIS_Status = factor(MIS_Status))
#sba <- separate(data = sba, col = ApprovalDate, into = c("day", "month", "year"), sep = "\\-")
sba %>% pull(MIS_Status) %>% table()
```

```
## .
##        CHGOFF  P I F
##   1997 157558 739609
```

Now, we check if the variable `ChgOffDate` can be used to measure the missing data. Since `ChgOffDate` corresponds to the date where the loan was charged off, we hypothesize that any columns with `ChgOffDate` but not `MIS_Status` can be filled in as "CHGOFF". However, we observe that not only does `ChgOffDate` not correspond necessarily to `MIS_Status` but that there are a decent number of loans that have a `ChgOffDate` and were paid in full later. Thus, we remove all the blank columns for `MIS_Status` as we cannot say anything about the final results of these loans.

```
sba <- sba %>% filter(MIS_Status != "")
sba$MIS_Status <- droplevels(sba$MIS_Status)
sba %>% pull(MIS_Status) %>% table()
```

```
## .
## CHGOFF  P I F
## 157558 739609
```

```
sba$MIS_Status <- ifelse(sba$MIS_Status == "P I F",1,0)
```

Only some variables will be used in this analysis. Excluded variables either contain too many categorical levels to be useful or are thought to not be useful.

```
## Considered variables
sba <- sba %>% select(State,Bank,NAICS,ApprovalFY,Term,NoEmp,NewExist,CreateJob,
                      RetainedJob,FranchiseCode,UrbanRural,LowDoc,RevLineCr,
                      DisbursementDate,DisbursementGross,BalanceGross,MIS_Status,
                      GrAppv,SBA_Appv)
```

We construct an `Industry` classification variable based on `NAICS`. For example, this variable will treat `NAICS` codes 31-33 as Manufacturing. We will also create a geographic regions variable to be used in place of states. The variable `DisbursementYear` will be created and used in place of `ApprovalFY`. This variable indicates the year in which businesses began receiving their loans.

```
NAIC <- data.frame(
  "Code" = c(11,21,22,23,31,32,33,42,44,45,48,49,51,52,53,54,55,
             56,61,62,71,72,81,92),
  "Industry"=c("Agriculture, Forestry, Fishing and Hunting",
               "Mining","Utilities","Construction",
               rep("Manufacturing",3),"Wholesale Trade",
               rep("Retail Trade",2),rep("Transportation and Warehousing",2),
               "Information","Finance and Insurance",
               "Real Estate Rental and Leasing",
               "Professional, Scientific, and Technical Services",
               "Management of Companies and Enterprises",
               "Administrative and Remediation Services",
               "Educational Services","Health Care and Social Assistance",
               "Arts, Entertainment, and Recreation",
               "Accommodation and Food Services",
               "Other Services (except Public Administration)",
               "Public Administration")
)

sba <- sba %>%
  mutate(NAICS = as.integer(NAICS / 10000)) %>%
  left_join(NAIC,by=c("NAICS"="Code")) %>%
  mutate(Industry = replace_na(Industry, "Unknown")) %>%
  filter(Industry != "Unknown") %>%
  mutate(Region = ifelse(State %in% c("CT", "MA", "DE", "ME", "NH",
                                      "NJ", "NY", "PA", "VT", "RI"), "NE", "O"),
       Region = ifelse(State %in% c("IL", "IN", "MI", "OH", "WI",
                                    "IA", "KS", "MN", "MO", "NE", "ND", "SD"),
                       "MW", Region),
       Region = ifelse(State %in% c("WA", "MT", "ID", "WY", "OR",
                                    "CA", "NV", "UT", "CO", "AZ",
                                    "NM"), "West", Region),
       Region = ifelse(Region == "O", "South", Region)) %>%
  mutate(Region = as.factor(Region))

sba$DisbursementDate = as.Date(sba$DisbursementDate, format="%d-%b-%y")
sba$DisbursementYear = year(sba$DisbursementDate)
```

We convert dollar amounts to numeric values using the `gsub` function.

```
sba <- sba %>%
  mutate(DisbursementGross = as.numeric(gsub("\\,", "",
    gsub("\\$", "", DisbursementGross)))) %>%
  mutate(BalanceGross = as.numeric(gsub("\\,", "",
                                        gsub("\\$", "", BalanceGross)))) %>%
  mutate(GrAppv = as.numeric(gsub("\\,", "", gsub("\\$", "", GrAppv)))) %>%
  mutate(SBA_Appv = as.numeric(gsub("\\,", "", gsub("\\$", "", SBA_Appv))))
```

We consider the number of employees on the log scale. Empty levels and levels of factor variables for which the meaning is unclear are removed. This is done for `UrbanRural`, `NewExist`, and `State`. Term length is converted to years as opposed to months, and entries with 0 term length are removed. We also create the variable `recession_2007` to model impacts of the Great Recession. All loans disbursed after 2011 are ignored because of the selection bias mentioned in the reference (https://www.tandfonline.com/doi/full/10.1080/10691898.2018.1434342).

A variable `p_GrAppv = GrAppv/SBA_Appv` is created. This variable is greater than 1 whenever the business receives more funding than what was approved by the SBA. This variable has the nice property that it is dimensionless in the physical sense, i.e. does not have units, and has the same meaning across years, different inflationary environments, and for large and small businesses alike. If you find the concept of modeling using "dimensionless" variables interesting then see this paper on work that bridges physical dimensional analysis and statistical modeling in the context of experimental design.

```
sba <- sba  %>%
  mutate(State = factor(State)) %>%
  mutate(log_noEmp = log(NoEmp + 1)) %>%
  filter(NewExist != 0) %>%
  filter(NoEmp > 0) %>%
  mutate(NewExist = as.factor(NewExist)) %>%
  mutate(UrbanRural = as.factor(UrbanRural)) %>%
  mutate(Industry = as.factor(Industry)) %>%
  mutate(Term = Term/12) %>%
  filter(Term > 0) %>%
  mutate(p_GrAppv = GrAppv/SBA_Appv) %>%
  mutate(DisbursementGross = log(DisbursementGross)) %>%
  mutate(recession_2007 = as.factor(ifelse(ApprovalFY >= 2007,1,0))) %>%
  filter(State != "") %>%
  filter(DisbursementYear < 2011) %>%
  filter(UrbanRural != 0) %>%
  filter(RevLineCr %in% c("Y","N",0)) %>%
  filter(LowDoc %in% c("Y","N")) %>%
  mutate(LowDoc = as.factor(LowDoc)) %>%
  mutate(RevLineCr = as.factor(RevLineCr))
sba$LowDoc <- droplevels(sba$LowDoc)
sba$RevLineCr <- droplevels(sba$RevLineCr)
sba$State <- droplevels(sba$State)
```

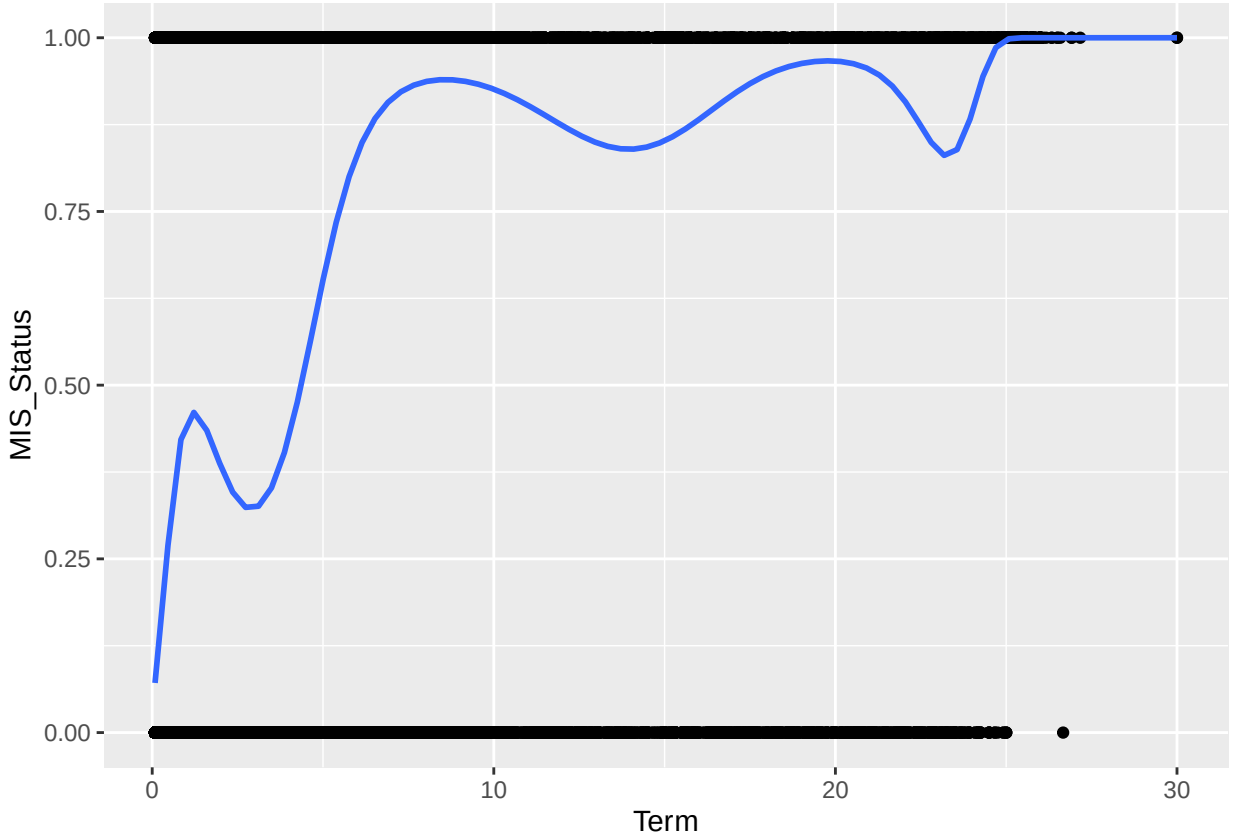
Data exploration

We will consider a subset of the data for exploratory purposes.

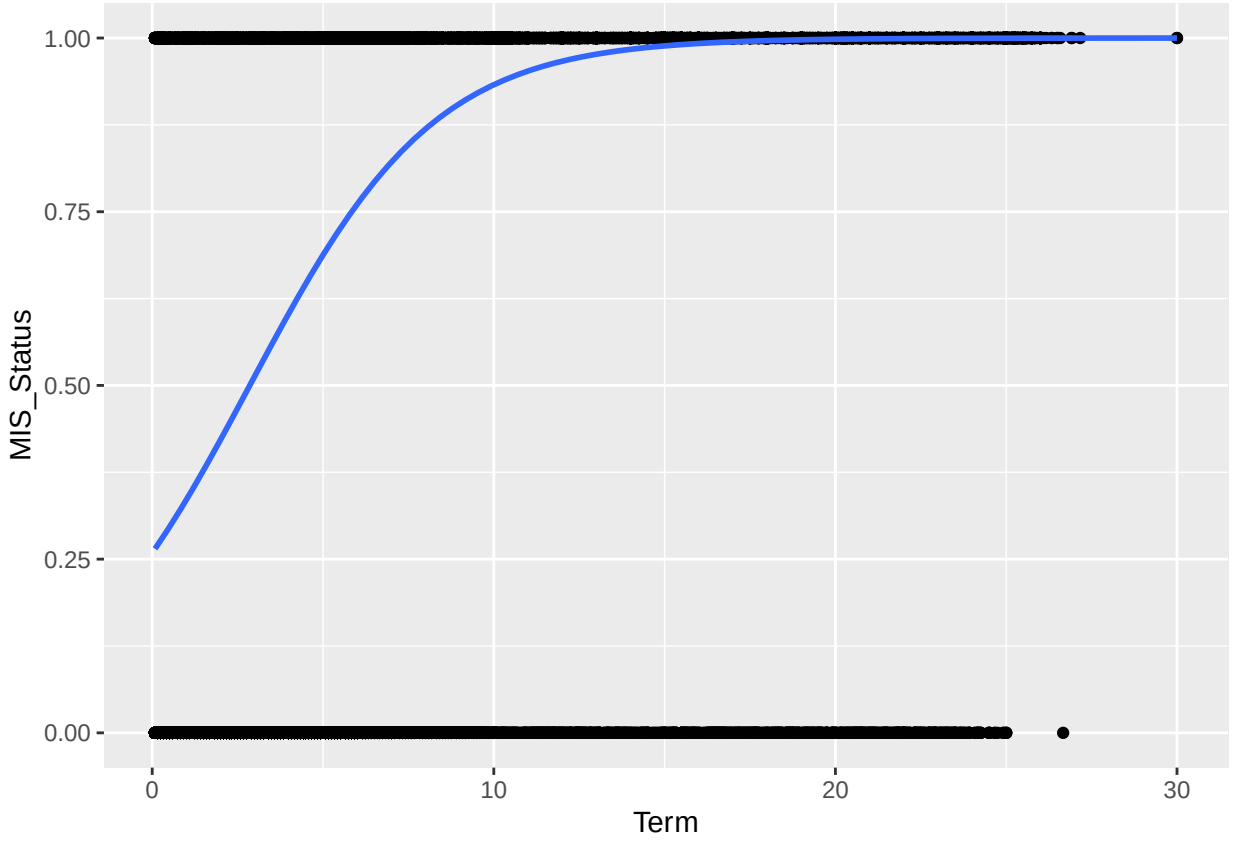
```
set.seed(13)
sba_test <- sba[sample(1:nrow(sba), size = 1e5, replace = FALSE), ]
```

The term length variable seems to be highly associated with the default rate. But this variable is irregular in that it has a large mass at certain discrete values.

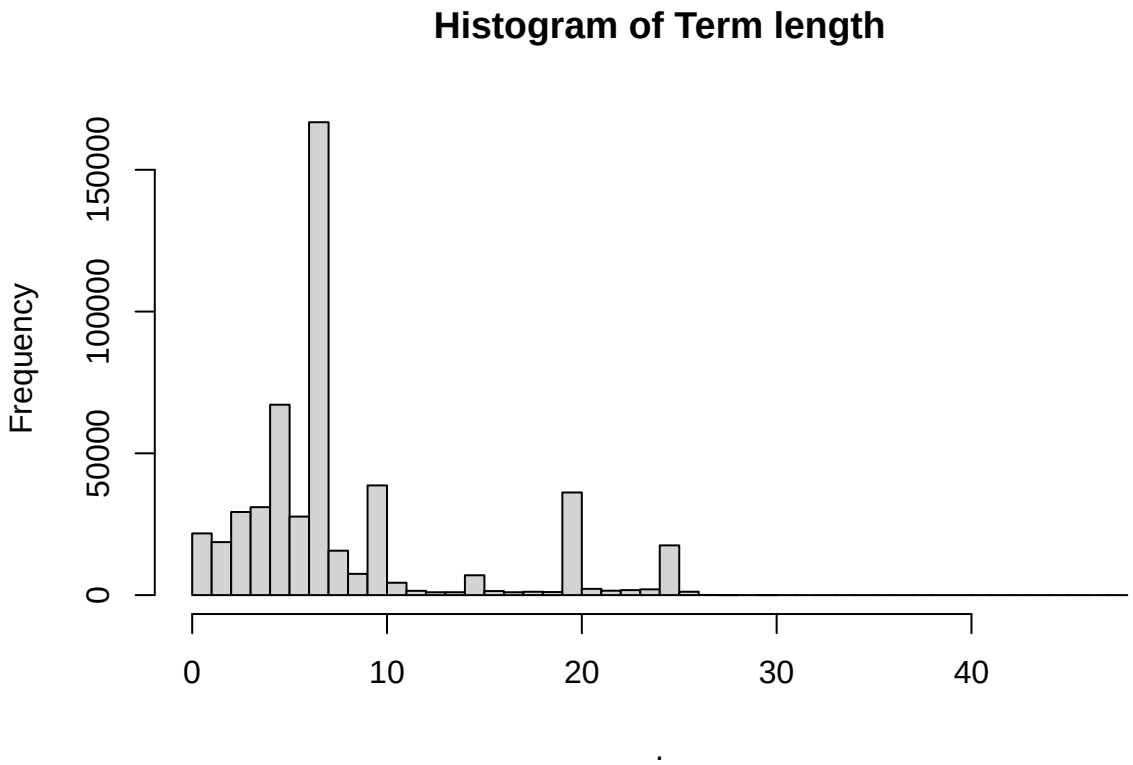
```
ggplot(sba_test, aes(x = Term, y = MIS_Status)) +
  geom_point() +
  geom_smooth(formula = y~poly(x, 10), method = "glm",
              method.args=list(family = "binomial"), se = FALSE)
```




```
ggplot(sba_test, aes(x = Term, y = MIS_Status)) +
  geom_point() +
  geom_smooth(method = "glm", method.args=list(family = "binomial"), se = FALSE)
```



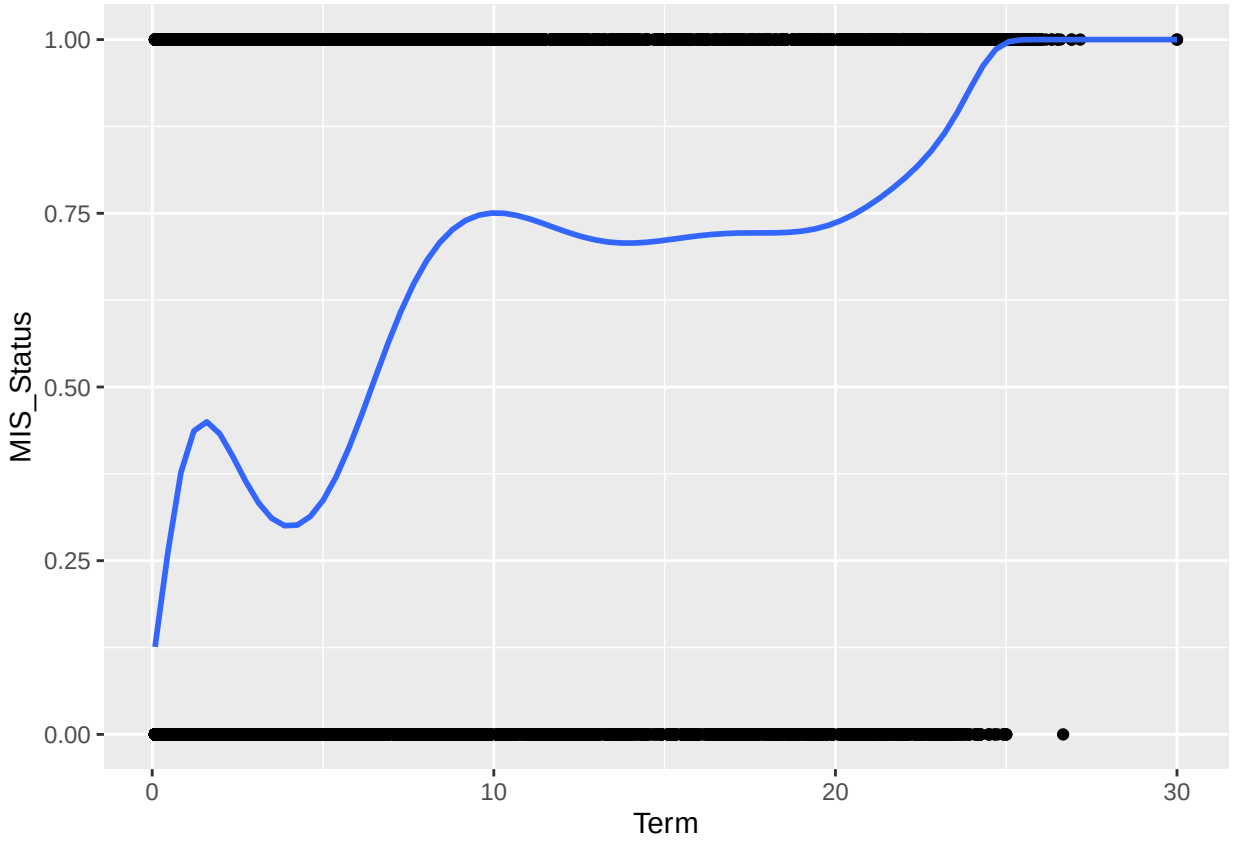

```
sba %>% pull(Term) %>% hist(., main = "Histogram of Term length", breaks = 50)
```



.

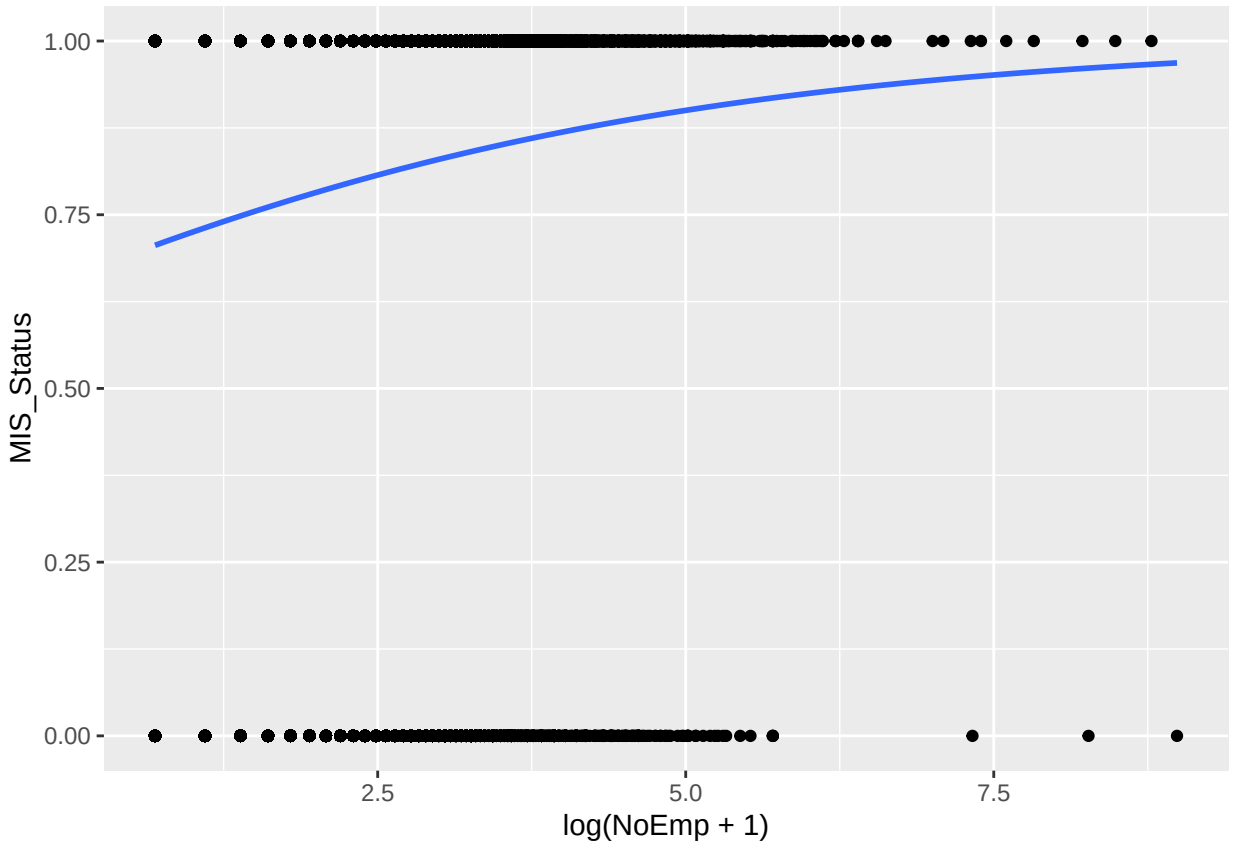
We still see that Term length is associated with the default rate after removing 5,7,10,15, and 20 year loans. It is interesting to note that the probability of paid in full decreases with the exclusion of these discrete term lengths. We will consider discrete effects for these categorical levels when we fit models.

```
ggplot(sba_test %>% filter(!Term %in% c(5,7,10,15,20)),
       aes(x = Term, y = MIS_Status)) +
  geom_point() +
    geom_smooth(formula = y~poly(x, 10), method = "glm",
                method.args=list(family = "binomial"), se = FALSE)
```



We see that the number of employees is also associated with the default rate. There is an interesting business with a large number of employees that defaulted. We will display some information on other large businesses which defaulted.

```
ggplot(sba_test, aes(x = log(NoEmp + 1), y = MIS_Status)) +
  geom_point() +
  geom_smooth(method = "glm", method.args=list(family = "binomial"), se = FALSE)
```




```
sba %>%
  filter(NoEmp > 2000, MIS_Status == 0) %>%
  select(DisbursementYear, Term, NoEmp, recession_2007, Industry)
```

```
##    DisbursementYear     Term NoEmp recession_2007
## 1:             2006 3.666667  8000              1
## 2:             2007 6.583333  3900              1
## 3:             2003 1.750000  7991              0
## 4:             2004 2.083333  4501              0
## 5:             2004 2.083333  4800              0
##                              Industry
## 1:      Transportation and Warehousing
## 2:                        Retail Trade
## 3: Arts, Entertainment, and Recreation
## 4:     Accommodation and Food Services
## 5:     Accommodation and Food Services
```

We see that some industries only have a few businesses receiving SBA loans. We will remove industries with fewer than 1000 businesses in the data set.

```
sba %>% pull(Industry) %>% table() %>% as.matrix()
```

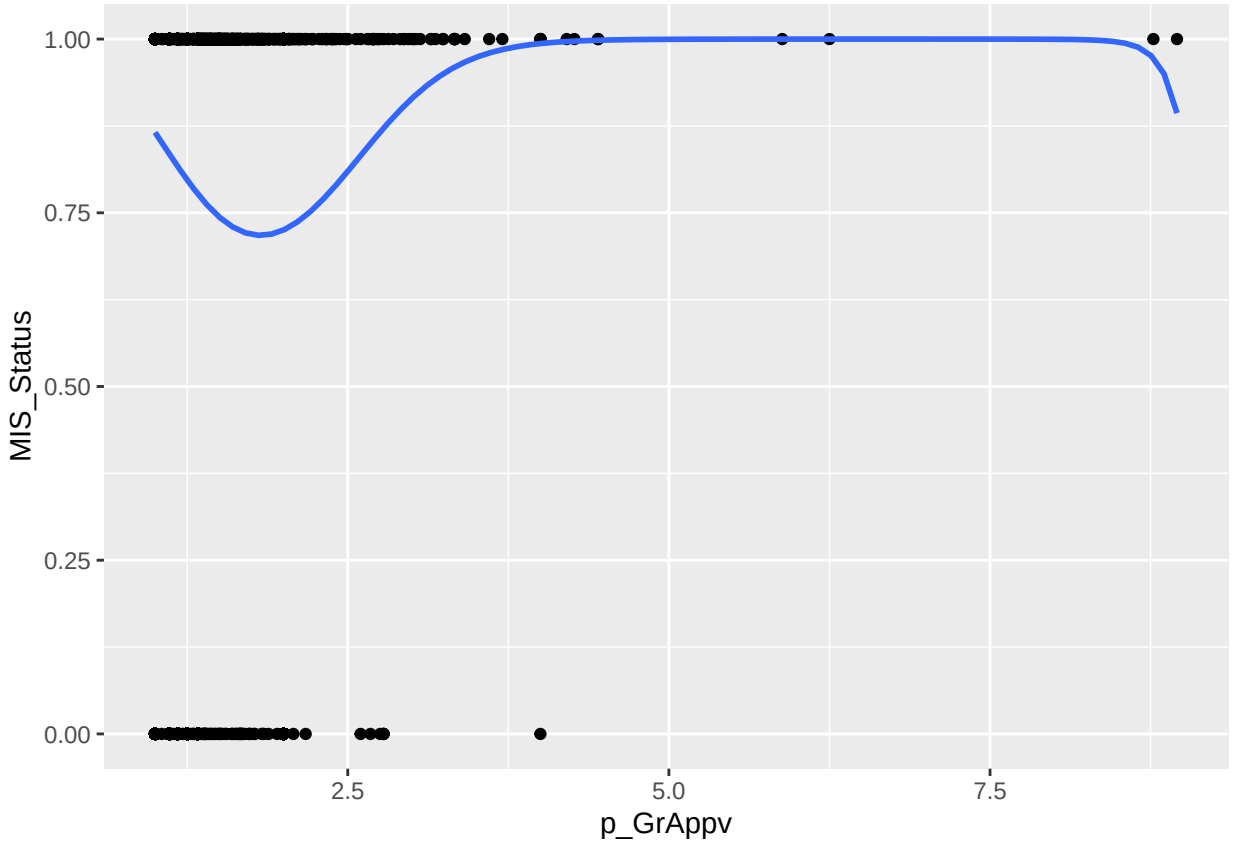
```
##                                                [,1]
## Accommodation and Food Services               58830
## Administrative and Remediation Services       26579
## Agriculture, Forestry, Fishing and Hunting     4103
## Arts, Entertainment, and Recreation           10306
## Construction                                  48012
## Educational Services                           5595
## Finance and Insurance                          8365
```

```
## Health Care and Social Assistance                35953
## Information                                       8685
## Management of Companies and Enterprises            158
## Manufacturing                                    45476
## Mining                                            1045
## Other Services (except Public Administration)    50540
## Professional, Scientific, and Technical Services 50002
## Public Administration                               91
## Real Estate Rental and Leasing                   10897
## Retail Trade                                     90852
## Transportation and Warehousing                   18660
## Utilities                                          412
## Wholesale Trade                                  31916
```

```
sba <- sba %>%
  filter(!Industry %in% c("Utilities","Public Administration",
                          "Management of Companies and Enterprises"))
```

We see that the default rate is associated with `p_GrAppv`. In particular, the default rate increases as the amount of funds dispersed relative to those approved by SBA increases.

```
ggplot(sba_test, aes(x = p_GrAppv, y = MIS_Status)) +
  geom_point() +
  geom_smooth(formula = y~poly(x, 3), method = "glm",
              method.args=list(family = "binomial"), se = FALSE)
```



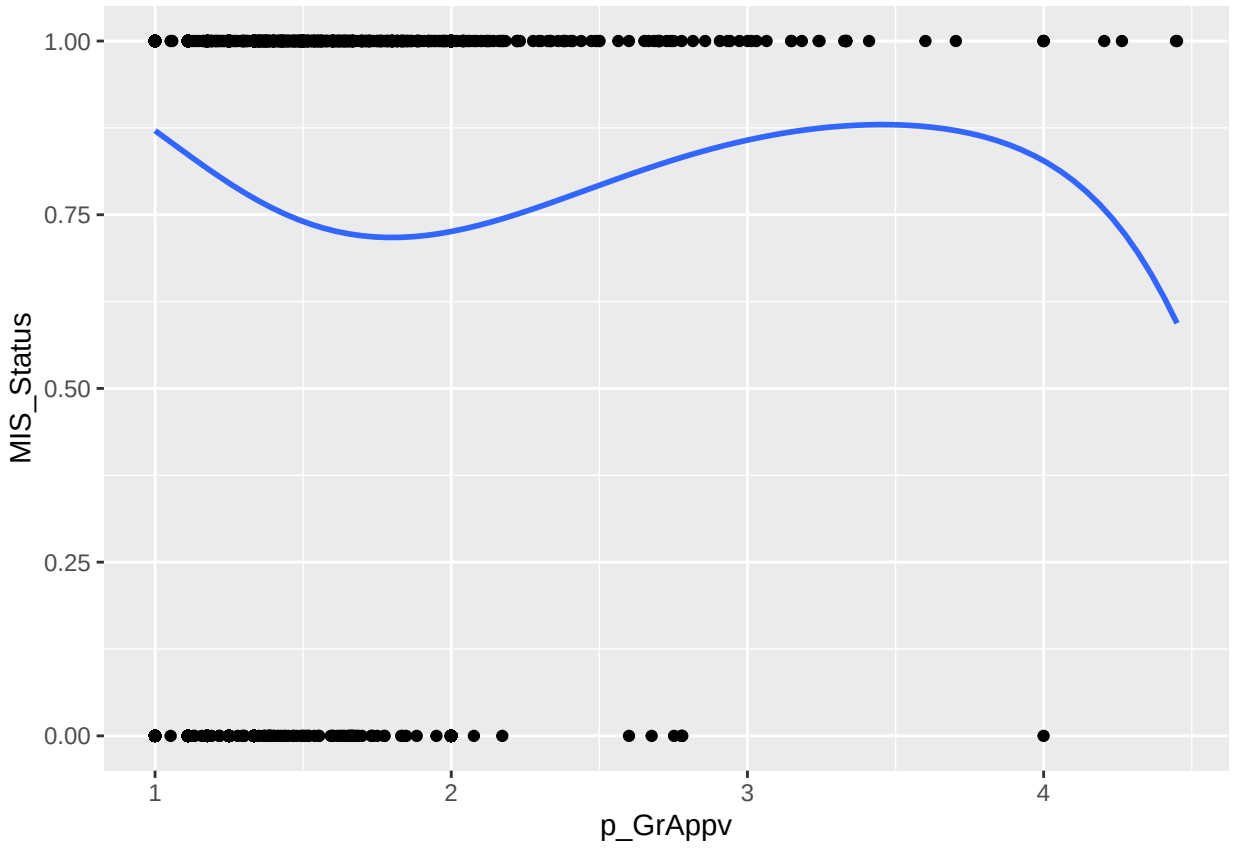
Outlying values of p_GrAppv are removed.

```
sba %>% select(DisbursementYear, Term, p_GrAppv, MIS_Status) %>%
  filter(p_GrAppv >= 5) %>%
  arrange(desc(p_GrAppv)) %>%
  as.data.frame()
```

```
##    DisbursementYear      Term  p_GrAppv MIS_Status
## 1              1999 25.000000 35.536649          1
## 2              2001 10.000000 20.000000          1
## 3              2004  7.000000 20.000000          1
## 4              2000 23.750000 17.857143          1
## 5              2006  6.000000 11.429852          1
## 6              1999 11.000000 11.420952          1
## 7              2004  7.916667 10.000000          1
## 8              2004 21.000000  9.090909          1
## 9              2001  6.000000  8.955759          1
## 10             2003 10.000000  8.771930          1
## 11             1999 25.000000  8.340301          1
## 12             2000 24.750000  7.258003          1
## 13             2003 23.000000  7.107321          1
## 14             2009  5.000000  6.666667          1
## 15             2005 25.000000  6.666667          1
## 16             2003 10.000000  6.377566          1
## 17             2001  1.000000  6.250000          1
## 18             2006 25.000000  5.882353          1
## 19             2007 10.250000  5.659313          1
## 20             2003  5.000000  5.639204          1
## 21             2000 10.500000  5.586614          1
## 22             2001 25.000000  5.482456          1
## 23             2001 20.000000  5.434515          1
## 24             2000 15.000000  5.405405          1
## 25             2000 25.000000  5.000000          1
## 26             2006  5.000000  5.000000          1
```

```
sba <- sba %>% filter(p_GrAppv <= 5)

ggplot(sba_test %>% filter(p_GrAppv <= 5), aes(x = p_GrAppv, y = MIS_Status)) +
  geom_point() +
  geom_smooth(formula = y~poly(x, 3), method = "glm",
              method.args=list(family = "binomial"), se = FALSE)
```



We will restrict attention to data entries in which the term length plus disbursement year is before 2015.

```
sba %>% pull(DisbursementYear) %>% table()
```

```
## .
##  1994  1995  1996  1997  1998  1999  2000  2001  2002  2003  2004  2005  2006
##    22   106    33   170    96  8373 21284 33725 42591 53606 66374 71191 74561
##  2007  2008  2009  2010
## 65007 33854 19627 15172
```

```
sba <- sba %>%
  filter(Term + DisbursementYear < 2015)
```

Finally, we will split data into training and testing sets.

```
set.seed(528)
idx = sample(nrow(sba), size = round(0.8*nrow(sba)))
train = sba[idx,]
test = sba[-idx,]
```

Modeling

We now fit a logistic regression model that considers several interactions and main-effect terms. Notably, we consider a linear effect for term length while also considering discrete effects for several common term lengths and short term loans. We consider post-2008 recession effects for several variables.

```
system.time(m <- glm(MIS_Status ~ Region*UrbanRural +
  p_GrAppv + I(p_GrAppv^2) + I(p_GrAppv^3) +
  RevLineCr + LowDoc +
  I(Term <= 2) + I(Term == 5) + I(Term == 7) + I(Term == 10) +
  Term + I(Term^2) + Term*p_GrAppv + log_noEmp + NewExist +
  Industry +
  recession_2007*(I(Term <= 2) + I(Term == 5) + I(Term == 7) +
                              Term + I(Term^2) + p_GrAppv + Region),
          data = train, family = "binomial"))
```

```
##    user  system elapsed
##   6.612   0.216   6.832
```

The summary table below reveals that most of these variables are statistically significant at any reasonable testing threshold. Justification for several of these covariates (and their effects on default rate) was investigated in the preceding section. It is interesting to note that the urban/rural effect interacts with geographical region, and that several covariates interact with the 2008 recession (encoded when funds were disbursed in the previous year). In particular the main-effect for term length is positively associated with the probability of paid in full before the recession, and is negatively associated after/during the recession.

```
summary(m)
```

```
##
## Call:
## glm(formula = MIS_Status ~ Region * UrbanRural + p_GrAppv + I(p_GrAppv^2) +
##     I(p_GrAppv^3) + RevLineCr + LowDoc + I(Term <= 2) + I(Term ==
##     5) + I(Term == 7) + I(Term == 10) + Term + I(Term^2) + Term *
##     p_GrAppv + log_noEmp + NewExist + Industry + recession_2007 *
##     (I(Term <= 2) + I(Term == 5) + I(Term == 7) + Term + I(Term^2) +
##         p_GrAppv + Region), family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.2896  -0.5733   0.0817   0.2520   2.5375
##
## Coefficients:
##                                                          Estimate Std. Error
## (Intercept)                                              7.715049   0.599629
## RegionNE                                                 0.669361   0.021681
## RegionSouth                                             -0.282668   0.021477
## RegionWest                                              -0.045568   0.021754
## UrbanRural2                                              0.448213   0.027733
## p_GrAppv                                               -16.115796   0.935539
## I(p_GrAppv^2)                                            7.166156   0.442399
## I(p_GrAppv^3)                                           -0.928764   0.063623
## RevLineCrN                                              -0.462677   0.017079
## RevLineCrY                                              -0.121855   0.017967
## LowDocY                                                  0.412622   0.026842
## I(Term <= 2)TRUE                                         1.025074   0.031113
## I(Term == 5)TRUE                                         3.476985   0.039543
## I(Term == 7)TRUE                                         4.684978   0.054851
## I(Term == 10)TRUE                                        4.661494   0.154915
## Term                                                    -0.020180   0.022242
## I(Term^2)                                                0.002271   0.001308
## log_noEmp                                                0.323247   0.007379
## NewExist2                                                0.206176   0.013620
## IndustryAdministrative and Remediation Services          0.459722   0.030444
## IndustryAgriculture, Forestry, Fishing and Hunting       1.002320   0.072209
## IndustryArts, Entertainment, and Recreation              0.171980   0.044658
## IndustryConstruction                                     0.471366   0.025459
## IndustryEducational Services                             0.321049   0.057924
## IndustryFinance and Insurance                           -0.003322   0.050400
```

```
## IndustryHealth Care and Social Assistance                  1.064809   0.031102
## IndustryInformation                                        0.308602   0.045540
## IndustryManufacturing                                      0.639710   0.026894
## IndustryMining                                             1.908177   0.142018
## IndustryOther Services (except Public Administration)      0.219517   0.026280
## IndustryProfessional, Scientific, and Technical Services   0.656787   0.026448
## IndustryReal Estate Rental and Leasing                    -0.177074   0.045315
## IndustryRetail Trade                                       0.177084   0.023003
## IndustryTransportation and Warehousing                     0.614456   0.032490
## IndustryWholesale Trade                                    0.560724   0.028641
## recession_20071                                            2.156179   0.127639
## RegionNE:UrbanRural2                                       0.063646   0.041872
## RegionSouth:UrbanRural2                                    0.087340   0.039099
## RegionWest:UrbanRural2                                     0.080953   0.043193
## p_GrAppv:Term                                              0.265707   0.008074
## I(Term <= 2)TRUE:recession_20071                          -0.025402   0.066108
## I(Term == 5)TRUE:recession_20071                           0.028806   0.058621
## I(Term == 7)TRUE:recession_20071                           0.506071   0.091884
## Term:recession_20071                                      -0.519431   0.041666
## I(Term^2):recession_20071                                  0.020295   0.004323
## p_GrAppv:recession_20071                                  -0.377530   0.041563
## RegionNE:recession_20071                                  -0.878314   0.037129
## RegionSouth:recession_20071                               -0.347456   0.035502
## RegionWest:recession_20071                                -1.010063   0.036267
##                                                          z value Pr(>|z|)
## (Intercept)                                               12.866  < 2e-16 ***
## RegionNE                                                  30.874  < 2e-16 ***
## RegionSouth                                              -13.161  < 2e-16 ***
## RegionWest                                                -2.095 0.036196 *
## UrbanRural2                                               16.162  < 2e-16 ***
## p_GrAppv                                                 -17.226  < 2e-16 ***
## I(p_GrAppv^2)                                             16.198  < 2e-16 ***
## I(p_GrAppv^3)                                            -14.598  < 2e-16 ***
## RevLineCrN                                               -27.090  < 2e-16 ***
## RevLineCrY                                                -6.782 1.18e-11 ***
## LowDocY                                                   15.372  < 2e-16 ***
## I(Term <= 2)TRUE                                          32.947  < 2e-16 ***
## I(Term == 5)TRUE                                          87.929  < 2e-16 ***
## I(Term == 7)TRUE                                          85.412  < 2e-16 ***
## I(Term == 10)TRUE                                         30.091  < 2e-16 ***
## Term                                                      -0.907 0.364251
## I(Term^2)                                                  1.736 0.082521 .
## log_noEmp                                                 43.804  < 2e-16 ***
## NewExist2                                                 15.138  < 2e-16 ***
## IndustryAdministrative and Remediation Services           15.101  < 2e-16 ***
## IndustryAgriculture, Forestry, Fishing and Hunting        13.881  < 2e-16 ***
## IndustryArts, Entertainment, and Recreation                3.851 0.000118 ***
## IndustryConstruction                                      18.515  < 2e-16 ***
## IndustryEducational Services                               5.543 2.98e-08 ***
## IndustryFinance and Insurance                             -0.066 0.947448
## IndustryHealth Care and Social Assistance                 34.236  < 2e-16 ***
## IndustryInformation                                        6.776 1.23e-11 ***
## IndustryManufacturing                                     23.786  < 2e-16 ***
## IndustryMining                                            13.436  < 2e-16 ***
## IndustryOther Services (except Public Administration)      8.353  < 2e-16 ***
## IndustryProfessional, Scientific, and Technical Services  24.833  < 2e-16 ***
## IndustryReal Estate Rental and Leasing                    -3.908 9.32e-05 ***
## IndustryRetail Trade                                       7.698 1.38e-14 ***
## IndustryTransportation and Warehousing                    18.912  < 2e-16 ***
## IndustryWholesale Trade                                   19.577  < 2e-16 ***
## recession_20071                                           16.893  < 2e-16 ***
## RegionNE:UrbanRural2                                       1.520 0.128510
## RegionSouth:UrbanRural2                                    2.234 0.025497 *
## RegionWest:UrbanRural2                                     1.874 0.060899 .
## p_GrAppv:Term                                             32.911  < 2e-16 ***
## I(Term <= 2)TRUE:recession_20071                          -0.384 0.700793
## I(Term == 5)TRUE:recession_20071                           0.491 0.623149
## I(Term == 7)TRUE:recession_20071                           5.508 3.64e-08 ***
```

```
## Term:recession_20071                                    -12.467  < 2e-16 ***
## I(Term^2):recession_20071                                 4.694 2.67e-06 ***
## p_GrAppv:recession_20071                                 -9.083  < 2e-16 ***
## RegionNE:recession_20071                                -23.656  < 2e-16 ***
## RegionSouth:recession_20071                              -9.787  < 2e-16 ***
## RegionWest:recession_20071                              -27.851  < 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: 370137  on 307361  degrees of freedom
## Residual deviance: 188546  on 307313  degrees of freedom
## AIC: 188644
##
## Number of Fisher Scoring iterations: 8
```

Results and validation

A likelihood ratio test suggests that our model fits the data better than a saturated model.

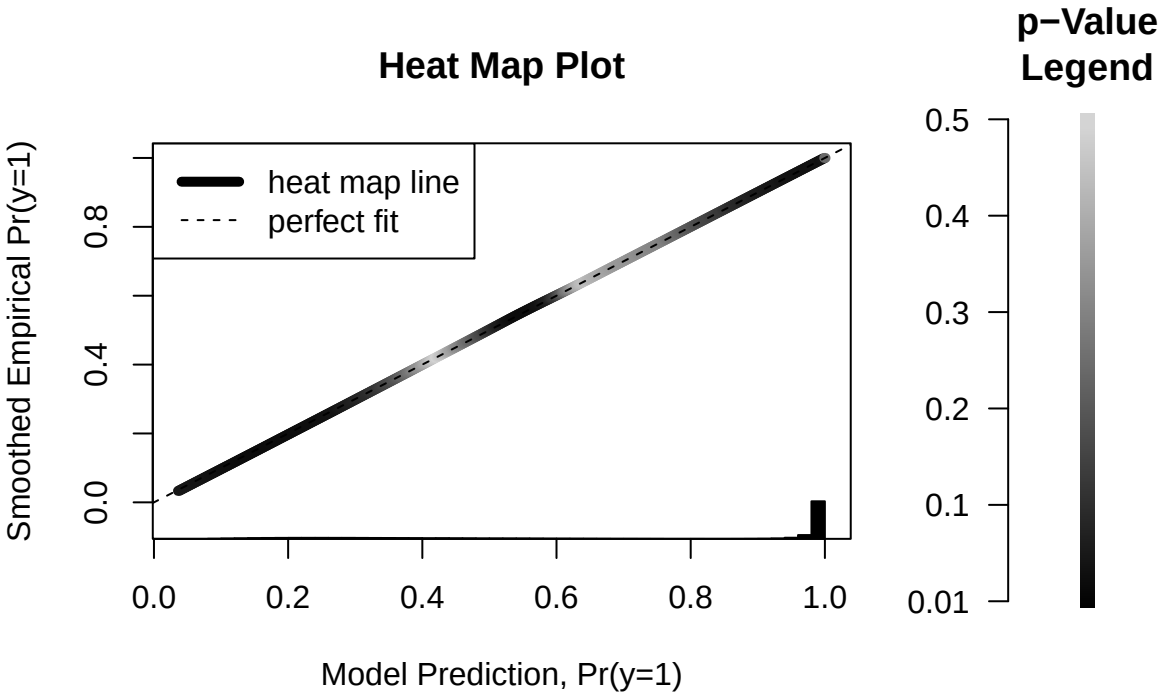
```
pchisq(deviance(m), m$df.residual, lower = FALSE)
```

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

The heatmap diagnostic in Esarey and Pierce (2012) suggests that our final model fits the data pretty well. The tests suggest a statistical lack of fit, but the heatmap plot does reveal any practical lack of fit. Model predictions is closely aligned with empirical success probabilities.

```
p_train <- predict(m, type = "response")
y_train <- train$MIS_Status
heatmap.fit(y_train, p_train)
```

```
## Collapsing large data set into bins based on predicted Pr(y)...
## Data collapsed into 1990 weighted bins.
##
## Calculating optimal loess bandwith...
## aicc Chosen Span =  0.9899459
##
## Generating Bootstrap Predictions...
##   |                                                                              |
```




```
##
##
## ****************************************
## 76.41088% of Observations have one-tailed p-value <= 0.1
## Expected Maximum = 20%
## ****************************************
```

We will now investigate classification using a confusion matrix on the test data where we encode a success when $\hat{p} > 0.5$, and a failure when $\hat{p} \leq 0.5$. From the output below, we see that classification accuracy of 0.8439 is much higher than the no information rate of 0.7081. Importantly, the sensitivity and specificity are, respectively, 0.9023 and 0.8198.

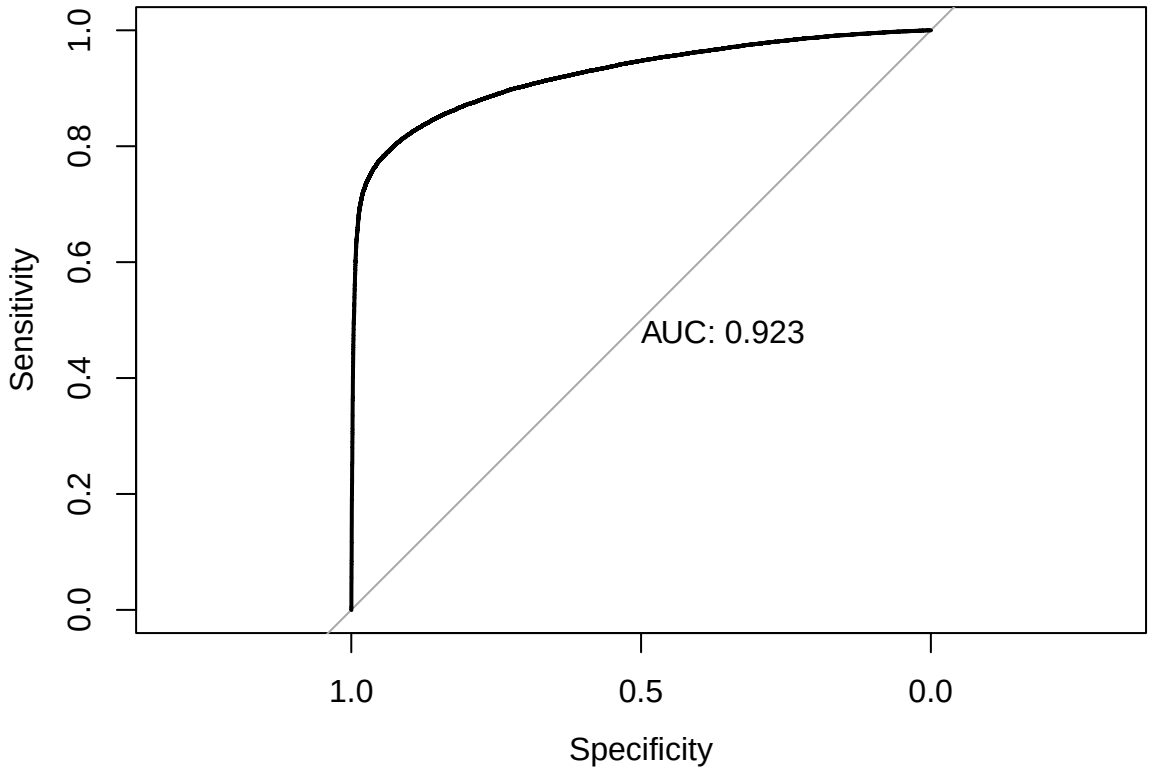
```
y_test <- test$MIS_Status
p_test <- predict(m, newdata = test)
conf_mat <- confusionMatrix(data=factor(as.numeric(p_test>0.5)),
                            reference = factor(y_test))
conf_mat
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction     0     1
##          0 20236  9806
##          1  2191 44608
##
##                Accuracy : 0.8439
##                  95% CI : (0.8413, 0.8464)
##     No Information Rate : 0.7081
```

```
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6566
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9023
##             Specificity : 0.8198
##          Pos Pred Value : 0.6736
##          Neg Pred Value : 0.9532
##              Prevalence : 0.2919
##          Detection Rate : 0.2633
##    Detection Prevalence : 0.3910
##       Balanced Accuracy : 0.8610
##
##        'Positive' Class : 0
##
```

A Receiver Operating Characteristic (ROC) curve is also provided. This ROC curve indicates that our logistic regression model is an adequate/good classifier as judged by area under the curve. Note that ROC curves were not covered in class, and you will not be penalized for not including an ROC curve.

```
sba_roc = roc(y_test ~ p_test, plot = TRUE, print.auc = TRUE)
```



Neural network We now compare our logistic regression model with a simple neural network. We first extract a model matrix for the main-effect terms used in our logistic regression model.

```
variables_train <- train %>% select(Term, NoEmp, log_noEmp, p_GrAppv,
                        Region, RevLineCr, LowDoc, NewExist, UrbanRural,
                        recession_2007)
modmat_train <- model.matrix(~., data = variables_train)
```

The neural network that we will consider will have 3 units in the hidden layer. We see that this neural network converged and it took a non-trivial amount of time to fit.

```
system.time(n1 <- nnet(x = modmat_train, y = y_train,
                       size = 3, maxit = 1e3, trace = FALSE))
```

```
##    user  system elapsed
##   5.288   0.033   5.331
```

```
n1$convergence
```

```
## [1] 0
```

This neural network had comparable, albeit lower, classification accuracy (0.8087) with our logistic regression model (0.8439). However, the neural network's sensitivity was relatively poor.

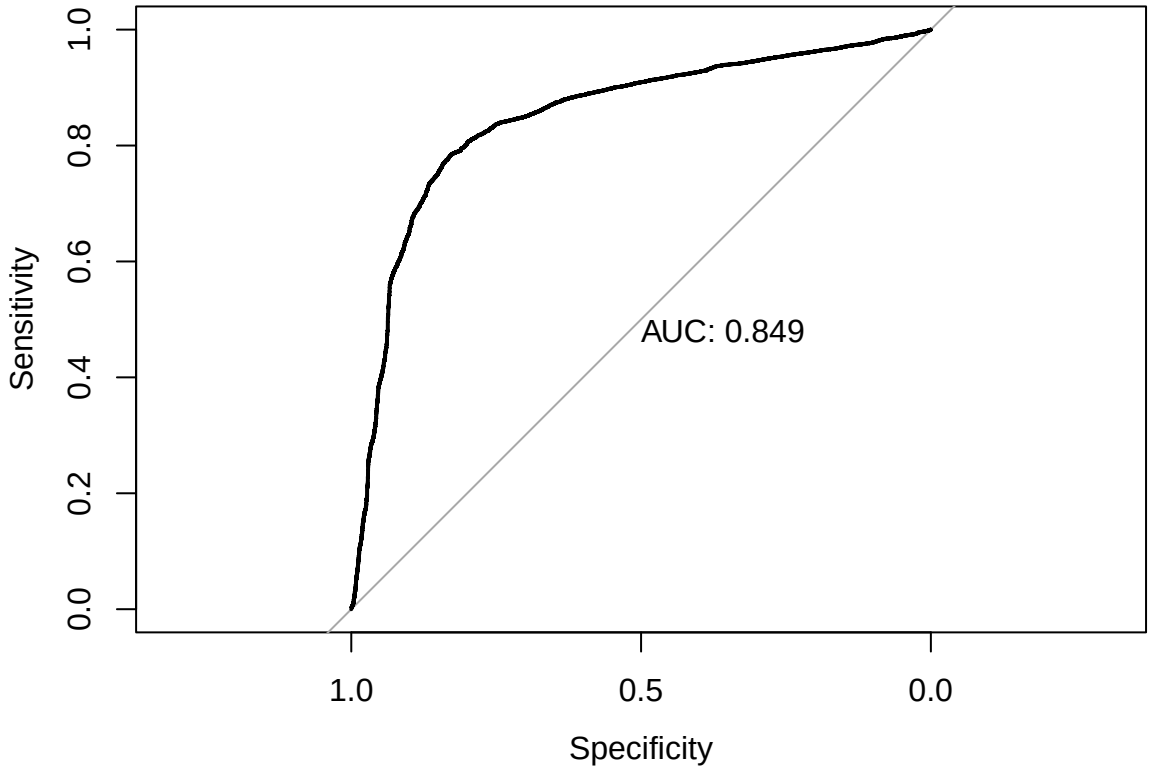
```
variables_test <- test %>% select(Term, NoEmp, log_noEmp, p_GrAppv,
                        Region, RevLineCr, LowDoc, NewExist,
                        UrbanRural, recession_2007)
modmat_test <- model.matrix(~., data = variables_test)
p_nnet <- predict(n1, newdata = modmat_test)
conf_mat_nnet <- confusionMatrix(
  data=factor(as.numeric(p_nnet>0.5)),
  reference = factor(test$MIS_Status))
conf_mat_nnet
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 16214  8486
##          1  6213 45928
##
##                Accuracy : 0.8087
##                  95% CI : (0.8059, 0.8115)
##     No Information Rate : 0.7081
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5506
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.7230
##             Specificity : 0.8440
##          Pos Pred Value : 0.6564
##          Neg Pred Value : 0.8808
##              Prevalence : 0.2919
```

```
##          Detection Rate : 0.2110
##    Detection Prevalence : 0.3214
##       Balanced Accuracy : 0.7835
##
##        'Positive' Class : 0
##
```

This ROC curve indicates that this neural network model is an adequate/good classifier and is comparable, albeit worse, than our logistic regression model as judged by area under the curve.

```
sba_roc_nnet = roc(y_test ~ p_nnet,
                   plot = TRUE, print.auc = TRUE)
```



Random forest We now compare our logistic regression model with a random forest.

```
system.time(rforest <- randomForest(y_train ~ .,
  data = modmat_train[, -1]))
```

```
##     user   system  elapsed
## 1669.594 1013.532 2686.258
```

```
p_rf <- predict(rforest, newdata = modmat_test[, -1])
```

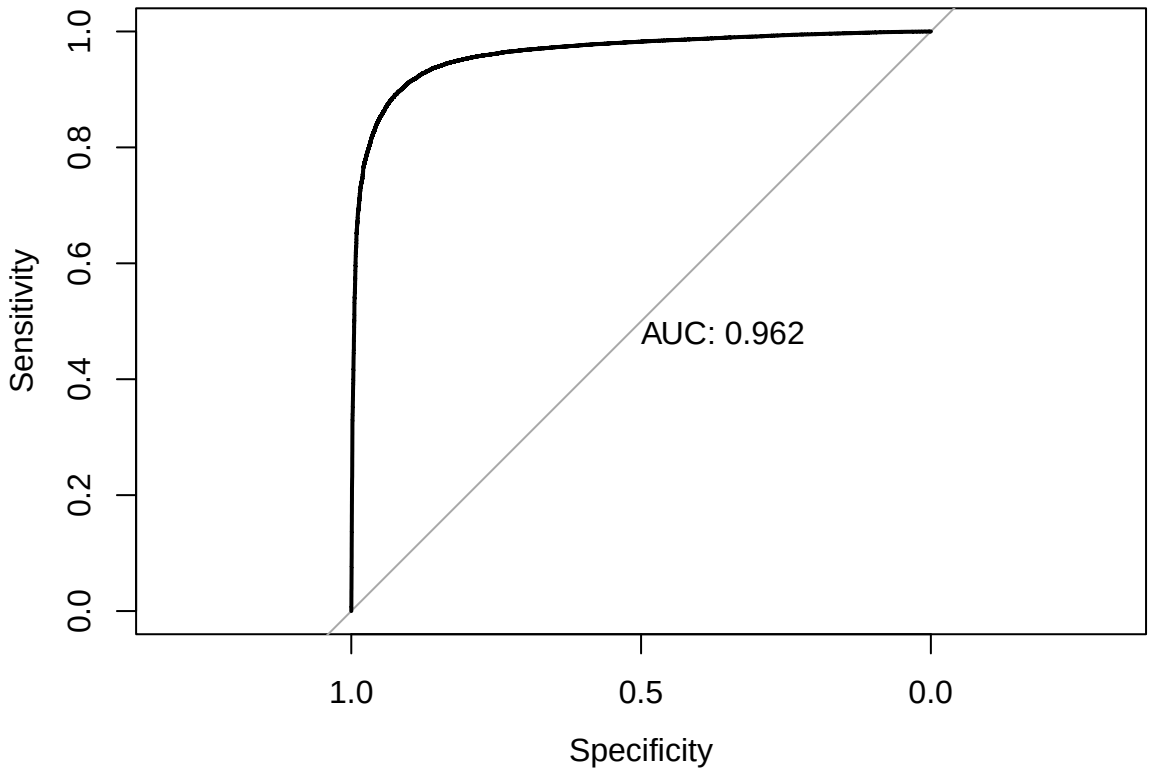
The random forest had superior classification accuracy (0.9139) with our logistic regression model (0.8439). It is interesting to note that our logistic regression model beat the random forest in terms of sensitivity (0.9023 vs 0.8583).

```
conf_mat_rf <- confusionMatrix(
  data=factor(as.numeric(p_rf>0.5)),
  reference = factor(y_test))
conf_mat_rf
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 19248  3434
##          1  3179 50980
##
##                Accuracy : 0.9139
##                  95% CI : (0.9119, 0.9159)
##     No Information Rate : 0.7081
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7925
##
##  Mcnemar's Test P-Value : 0.001787
##
##             Sensitivity : 0.8583
##             Specificity : 0.9369
##          Pos Pred Value : 0.8486
##          Neg Pred Value : 0.9413
##              Prevalence : 0.2919
##          Detection Rate : 0.2505
##    Detection Prevalence : 0.2952
##       Balanced Accuracy : 0.8976
##
##        'Positive' Class : 0
##
```

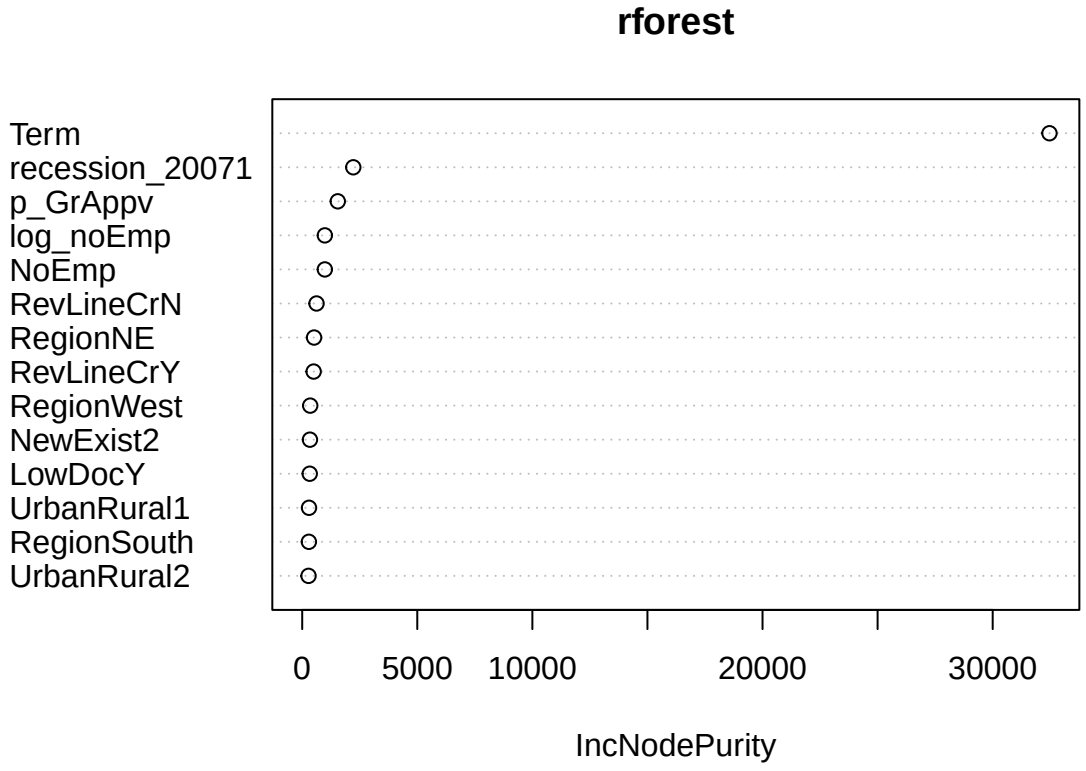
This ROC curve indicates that the random forest is an good classifier and is comparable, albeit better, than our logistic regression model as judged by area under the curve.

```
sba_roc_rf = roc(y_test ~ p_rf,
                  plot = TRUE, print.auc = TRUE)
```



A variable importance plot reveals that term length is far and away the most relevant predictor.

```
varImpPlot(rforest, sort = TRUE)
```



Concluding remarks

1. Our logistic regression model included highly relevant predictors, is better fitting than a saturated model, passes model diagnostics, and exhibits good in-sample classification accuracy. This model does not include year effects or covariates that are recorded after loan funds are disbursed, and can therefore be used to assess the chances that a loan will be paid in full after the loan term length is designated.

2. The effect that term length has on the probability that a loan is paid in full is interesting as it exhibits a linear continuous type effect punctuated by jumps for specific frequently occurring term lengths as well as a short term length effect. It could be that specific term lengths such as 5,7, and 10 year loans are a part of standard banking practices, and that other less occurring term lengths reflect different banking practices. In any event, these discrete term length variables are relevant for modeling. A follow up analysis in the Appendix demonstrates that the removal of these effects yields poor model fit.

3. We compared our logistic regression model with a simple neural network, and this comparison was favorable for our logistic regression. However, it should be noted that not much attention was given toward properly tuning the neural network. So it may be a bit premature to conclude that logistic regression modeling is better than a neural network based approach. That being said, the discrete and continuous nature of the term length variable may be difficult for neural networks.

4. We also compared our logistic regression model with a random forest, and this comparison was largely favorable for the random forest in terms of classification accuracy. Interestingly, the random forest did not win across the board in classification accuracy. The logistic regression model wins in sensitivity. Perhaps our logistic regression model can be improved by better modeling regions with higher probabilities of default. In the Appendix we go a different direction and explore an ensemble based approach to improve our logistic regression model by using the random forest predicted probabilities as a model input. This model is better on balance than both our original logistic regression model and the original random forest model, but it still loses in sensitivity.

5. Several potentially interesting variables were excluded. We did not consider inflation-adjusted disbursement gross, state effects, or recession effects for industries in this analysis. We also did not include the franchise code variable as we thought that new vs established business was more relevant.

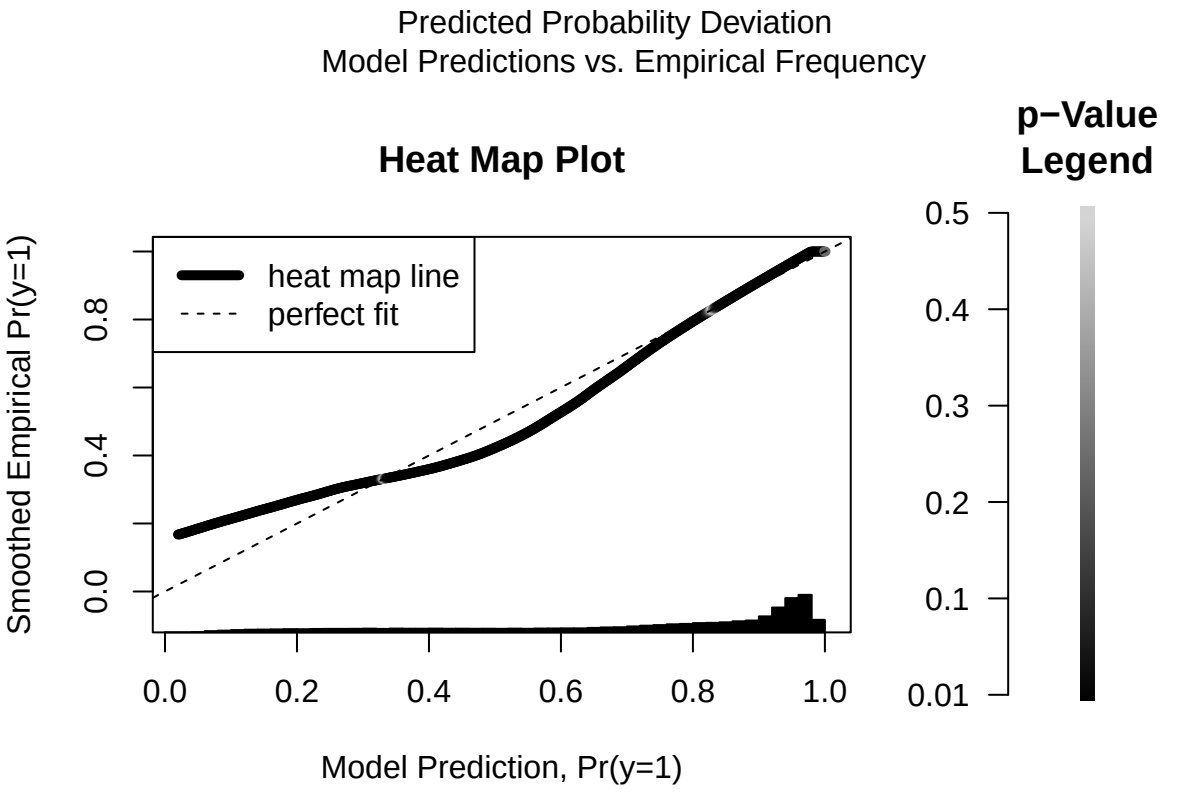
Appendix (removal of discrete term effects)

Here we refit our logistic regression model without the discrete Term effects. This model exhibits severe lack of fit.

```
m2 <- glm(MIS_Status ~ Region + p_GrAppv + I(p_GrAppv^2) + I(p_GrAppv^3) +
           RevLineCr + LowDoc + Term + Term*p_GrAppv + log_noEmp +
            NewExist + UrbanRural + Region*UrbanRural + Industry +
            recession_2007*(Term + p_GrAppv + Region),
          data = train, family = "binomial")

p2 <- predict(m2, type = "response")
heatmap.fit(y_train, p2)
```

```
## Collapsing large data set into bins based on predicted Pr(y)...
## Data collapsed into 1989 weighted bins.
##
## Calculating optimal loess bandwith...
## aicc Chosen Span =  0.517219
##
## Generating Bootstrap Predictions...
##   |                                                                      |
```




```
##
##
## ****************************************
## 97.80942% of Observations have one-tailed p-value <= 0.1
## Expected Maximum = 20%
## ****************************************
```

Appendix (ensemble model based approach)

We fit the same logistic regression model as before with the random forest estimated success probabilities as an additional covariate.

```
p_rf_train <- predict(rforest)
train_super <- train
train_super$rf <-  p_rf_train
system.time(m_super <- glm(MIS_Status ~ rf + Region*UrbanRural +
                        p_GrAppv + I(p_GrAppv^2) + I(p_GrAppv^3) +
                        RevLineCr + LowDoc +
                        I(Term <= 2) + I(Term == 5) + I(Term == 7) + I(Term == 10) +
                        Term + I(Term^2) + Term*p_GrAppv + log_noEmp + NewExist +
                        Industry +
                        recession_2007*(I(Term <= 2) + I(Term == 5) + I(Term == 7) +
                                        Term + I(Term^2) + p_GrAppv + Region),
                    data = train_super, family = "binomial"))
```

```
##    user  system elapsed
##   6.337   0.221   6.561
```

This model has the highest classification accuracy and its ROC curve has the highest area under the curve.

```
test_super <- test
test_super$rf <- p_rf
p_super_test <- predict(m_super, newdata = test_super, type="response")

conf_mat_super <- confusionMatrix(
  data=factor(as.numeric(p_super_test>0.5)),
  reference = factor(y_test))
conf_mat_super
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 19421  3326
##          1  3006 51088
##
##                Accuracy : 0.9176
##                  95% CI : (0.9156, 0.9195)
##     No Information Rate : 0.7081
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8015
##
##  Mcnemar's Test P-Value : 6.101e-05
##
##             Sensitivity : 0.8660
##             Specificity : 0.9389
##          Pos Pred Value : 0.8538
##          Neg Pred Value : 0.9444
##              Prevalence : 0.2919
```

```
##          Detection Rate : 0.2527
##    Detection Prevalence : 0.2960
##       Balanced Accuracy : 0.9024
##
##        'Positive' Class : 0
##
```

```
sba_roc_rf = roc(y_test ~ p_super_test,
                 plot = TRUE, print.auc = TRUE)
```