

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)
```

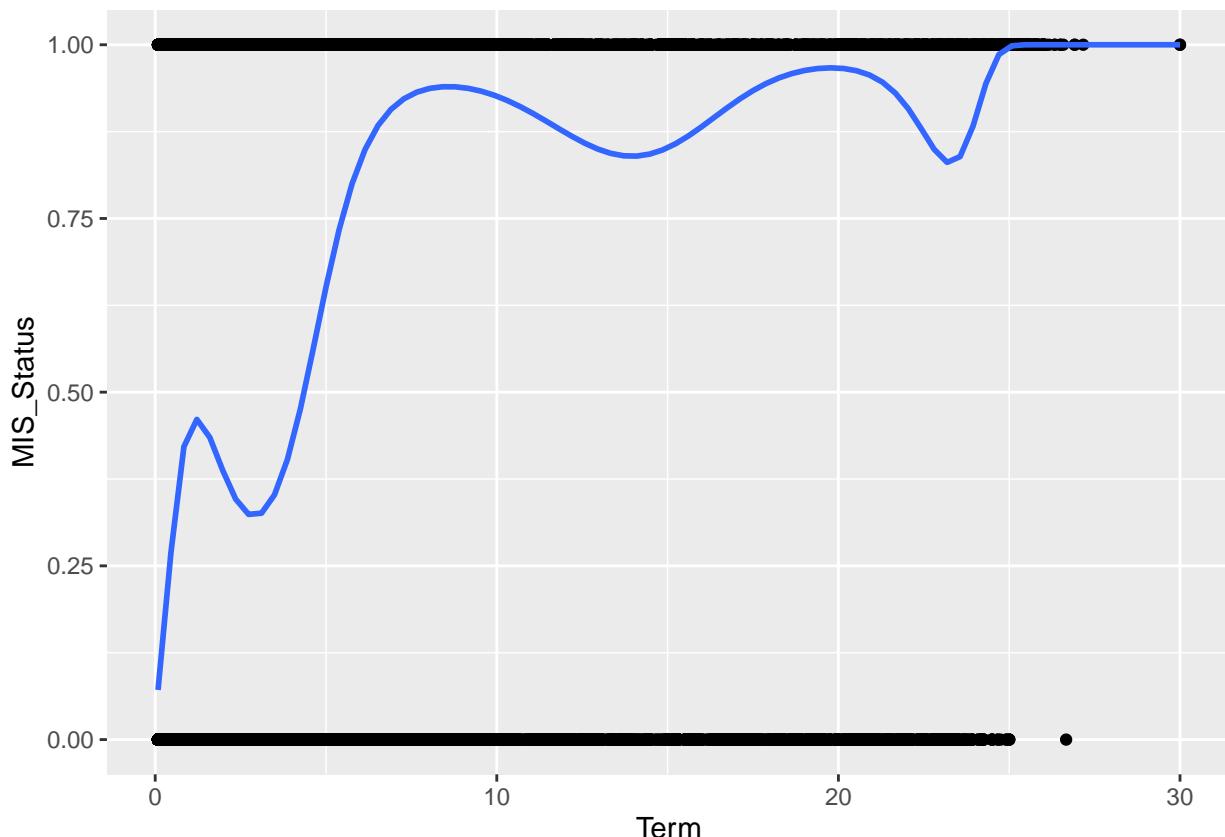
Data exploration

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

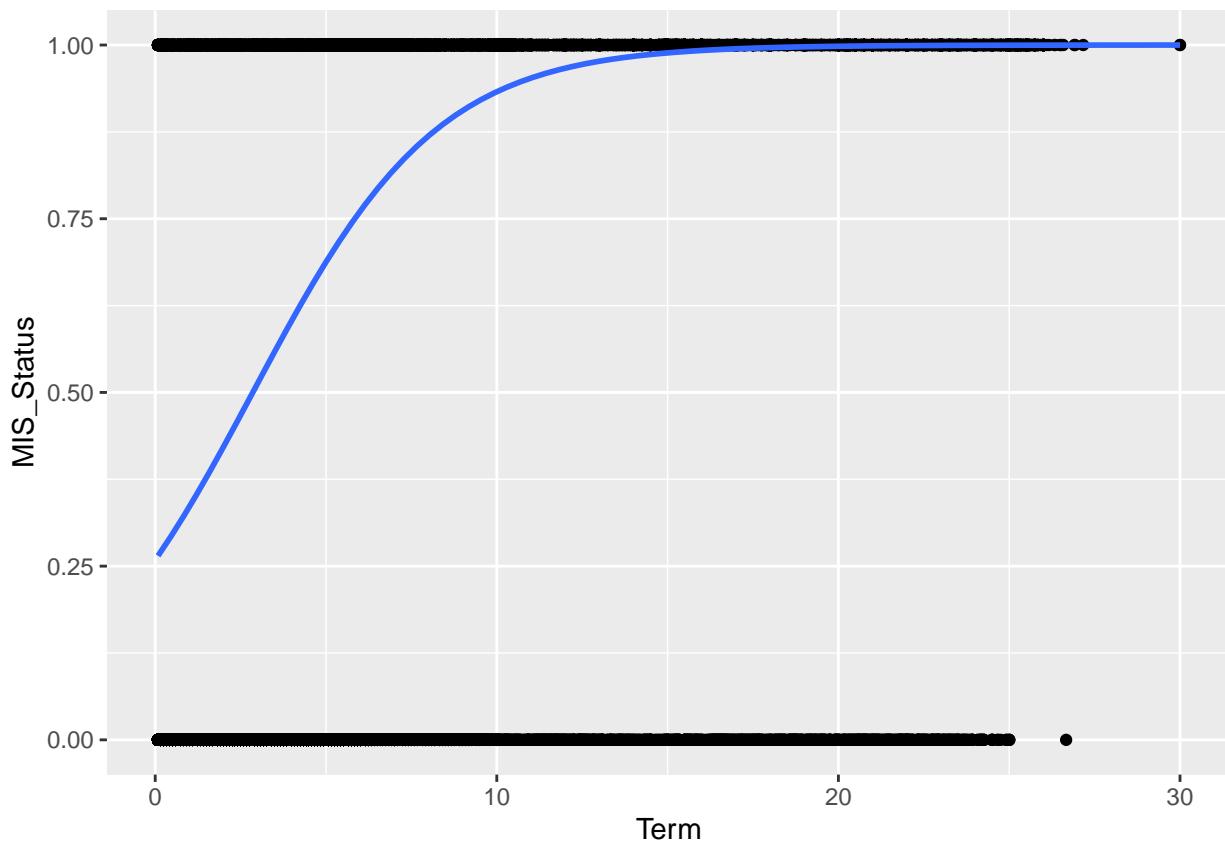
```
set.seed(13)
sba_test <- sba[sample(1:nrow(sba), size = 1e5, replace = FALSE), ]
```

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.

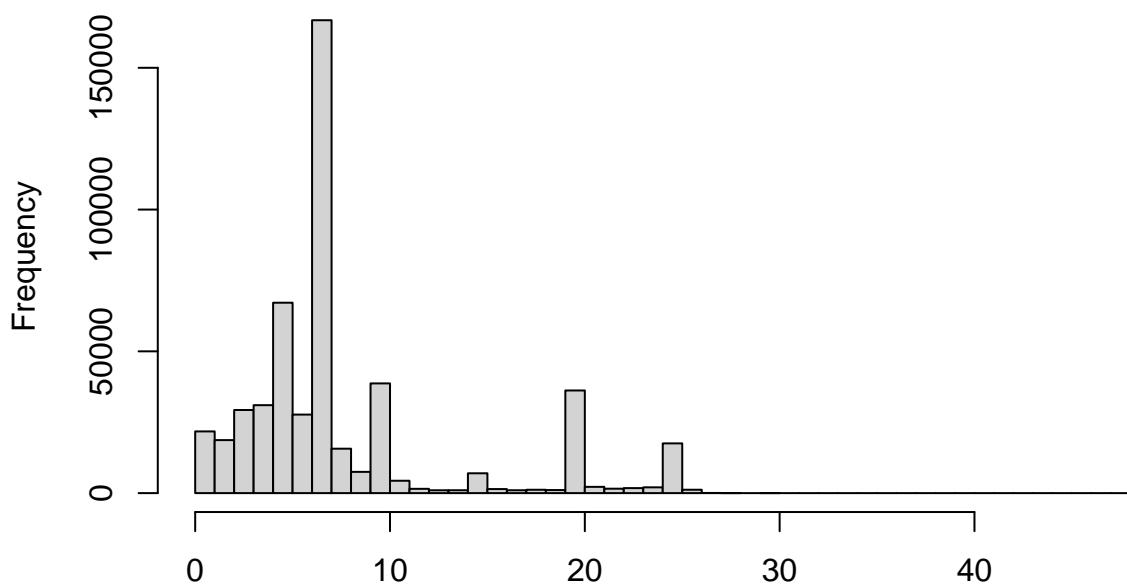
```
ggplot(sba_test, aes(x = Term, y = MIS_Status)) +
  geom_point() +
  geom_smooth(formula = y~poly(x, 10), method = "glm",
              method.args=list(family = "binomial"), se = FALSE)
```



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

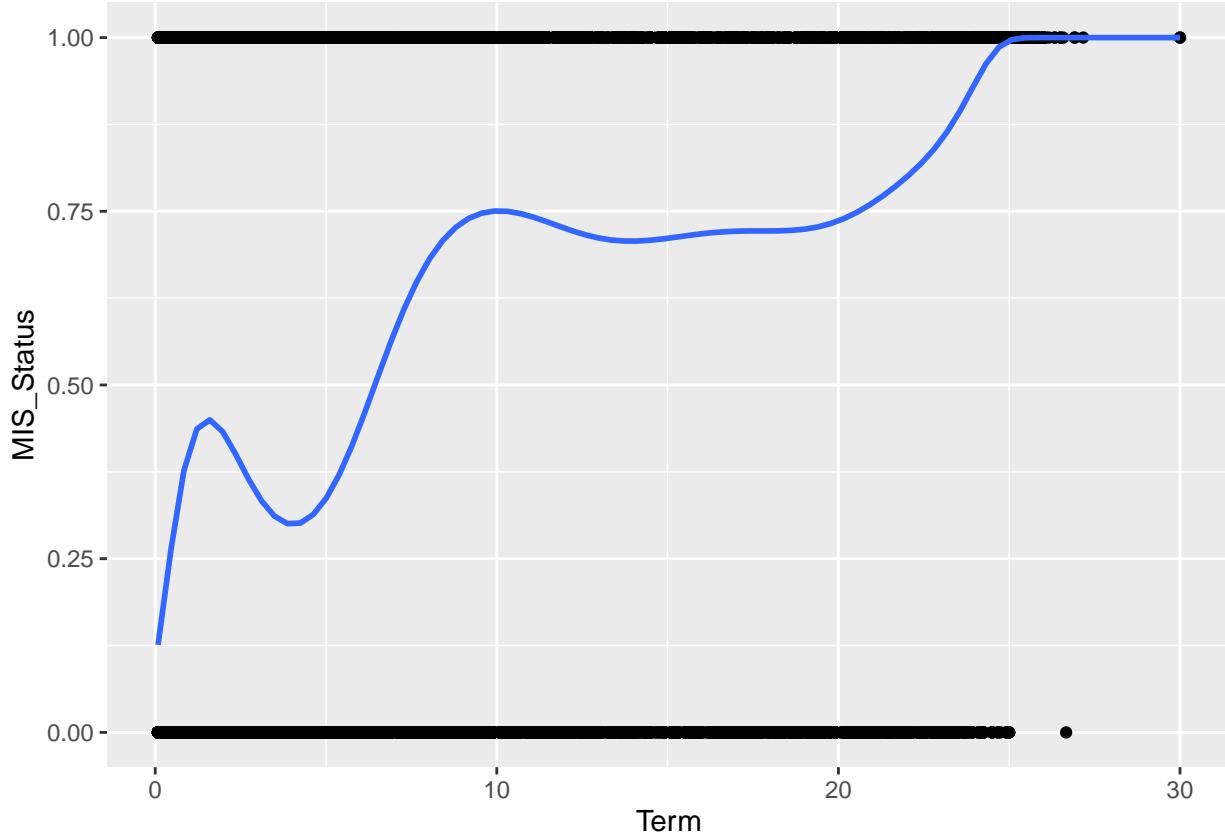


Histogram of Term length



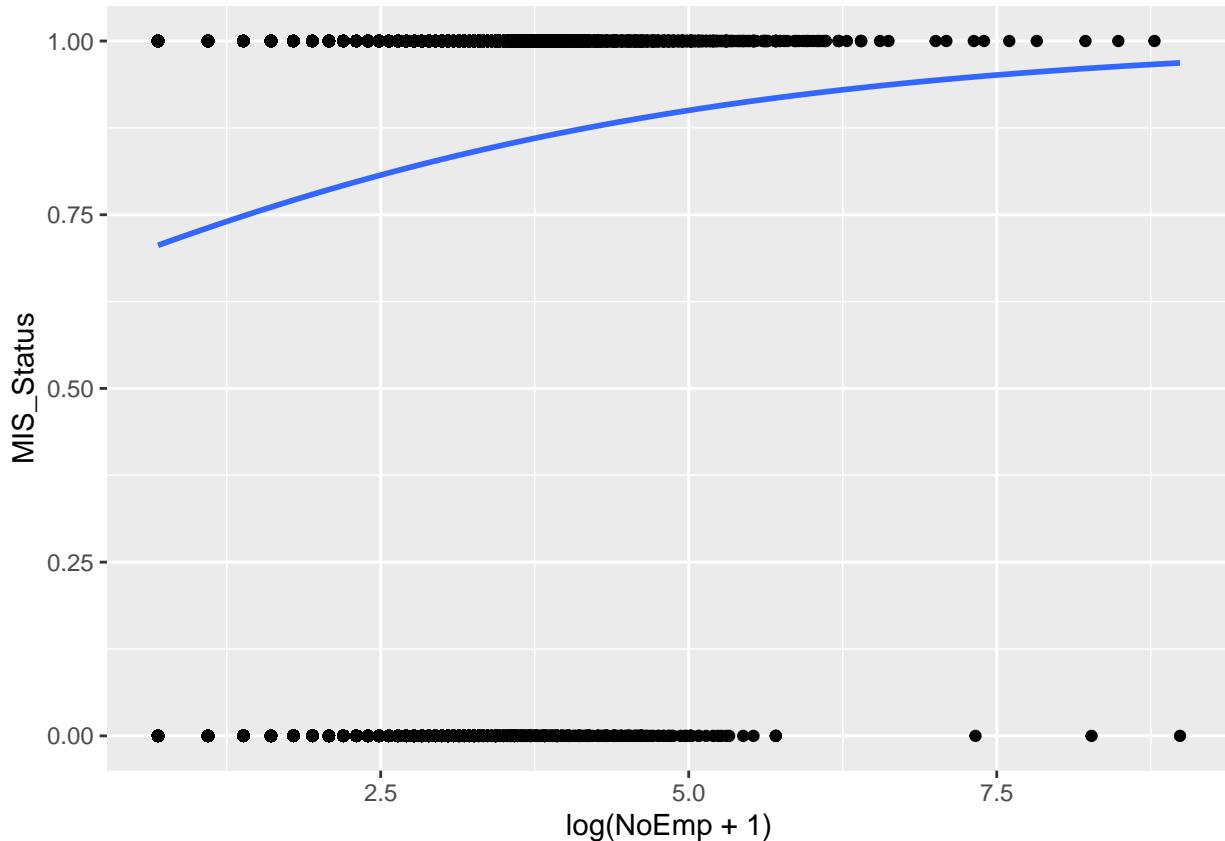
We still see that Term length is associated with the default rate after removing 5,7,10,15, and 20 year loans. It is interesting to note that the probability of paid in full decreases with the exclusion of these discrete term lengths. We will consider discrete effects for these categorical levels when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

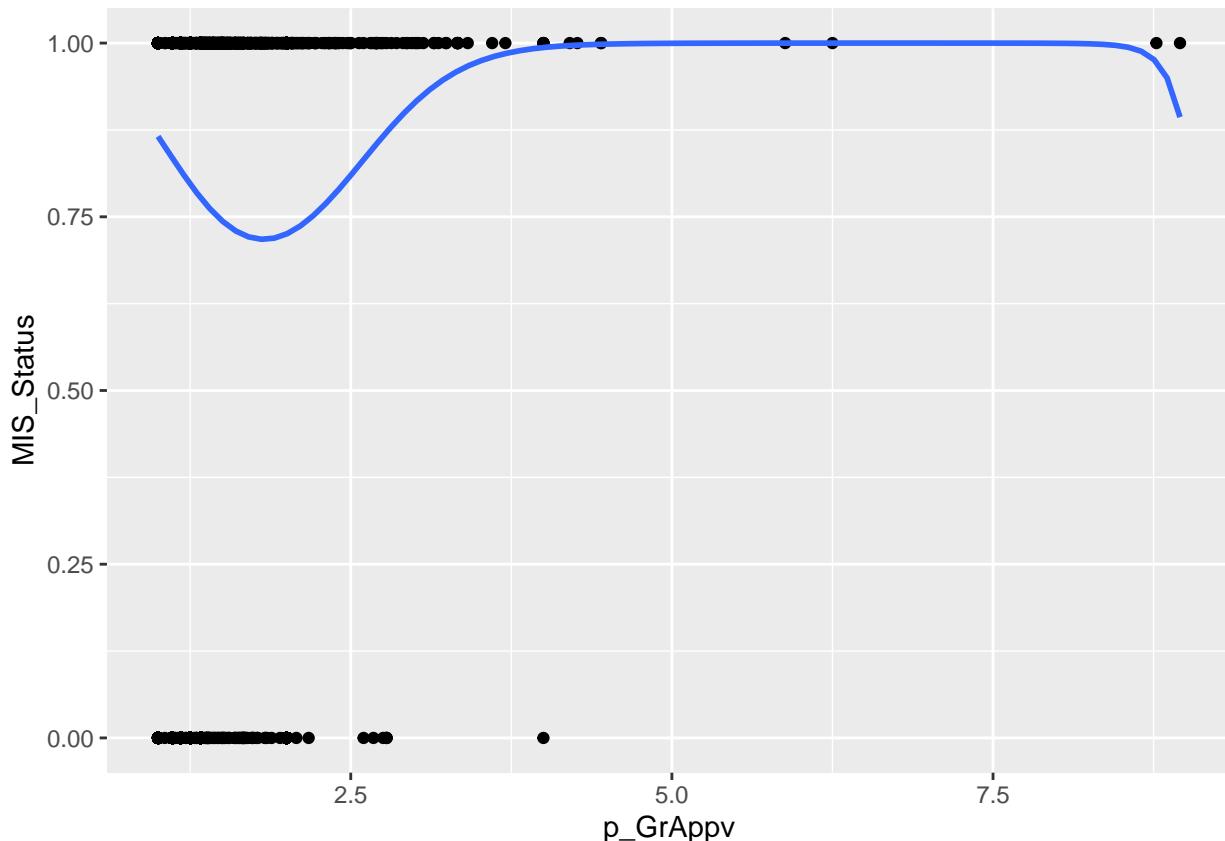
```

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)

```



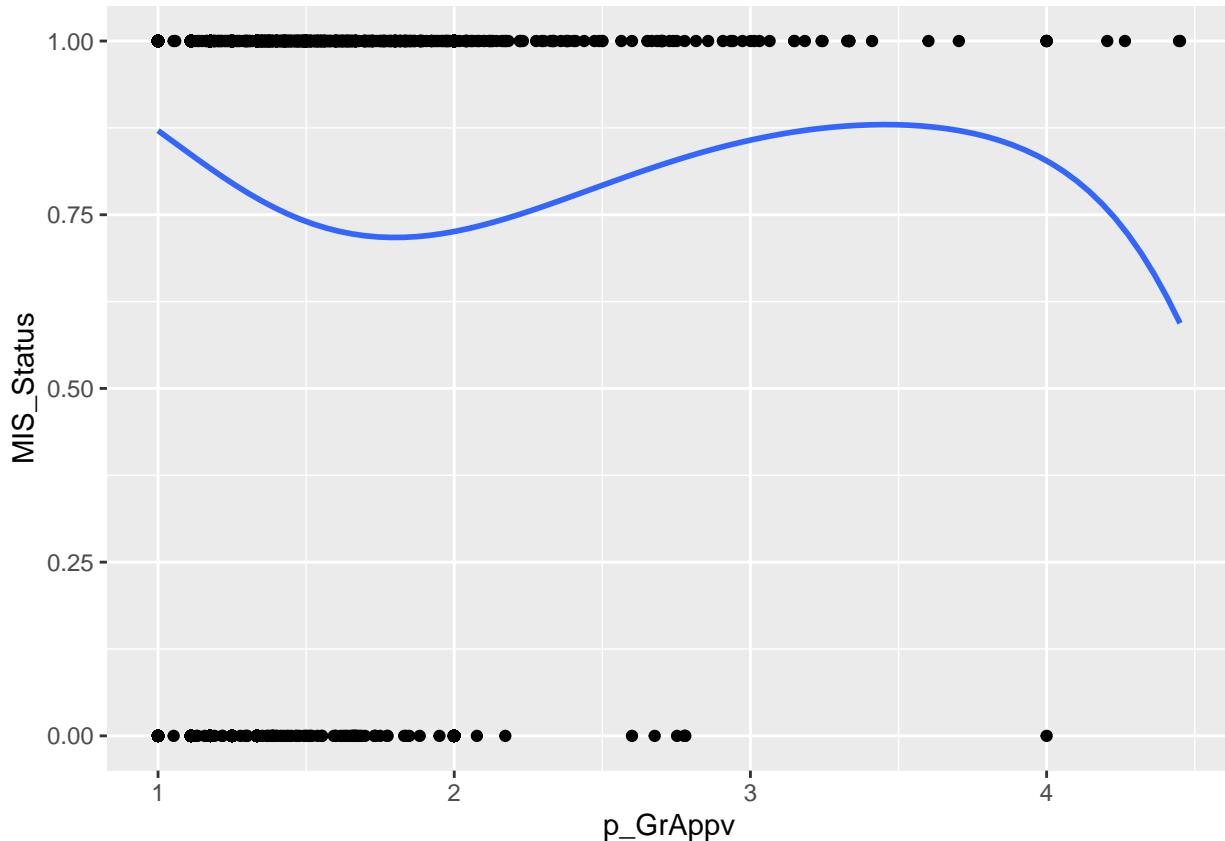
Outlying values of p_GrAppv are removed.

```
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))*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
##     8.761    0.267   9.029

```

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)) * 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
## -4.0970   -0.5650    0.0835    0.2546    2.6027
##
## Coefficients:
##                                         Estimate Std. Error
## (Intercept)                         5.116e+00  6.460e-01
## RegionNE                            6.559e-01  2.167e-02
## RegionSouth                          -2.826e-01 2.153e-02
## RegionWest                           -5.222e-02 2.177e-02
## UrbanRural2                          4.525e-01  2.784e-02
## p_GrAppv                            -1.445e+01 9.633e-01
## I(p_GrAppv^2)                        6.947e+00  4.535e-01
## I(p_GrAppv^3)                        -8.941e-01 6.559e-02
## RevLineCrN                           -4.717e-01 1.718e-02
## RevLineCrY                           -1.361e-01 1.799e-02
## LowDocY                             4.457e-01  2.754e-02
## I(Term <= 2)TRUE                   3.401e+00  1.740e-01
## I(Term == 5)TRUE                   4.195e+00  1.813e-01
## I(Term == 7)TRUE                   4.847e+00  2.403e-01
## I(Term == 10)TRUE                  8.690e+00  8.914e-01
## Term                                 4.880e-01  7.726e-02
## I(Term^2)                            -1.607e-02 6.654e-03
## log_noEmp                           3.224e-01  7.395e-03
## NewExist2                           2.116e-01  1.367e-02
## IndustryAdministrative and Remediation Services 4.652e-01  3.051e-02
## IndustryAgriculture, Forestry, Fishing and Hunting 9.823e-01  7.262e-02
## IndustryArts, Entertainment, and Recreation 1.752e-01  4.479e-02
## IndustryConstruction                4.725e-01  2.550e-02
## IndustryEducational Services        3.292e-01  5.806e-02
## IndustryFinance and Insurance      5.824e-04  5.034e-02

```

| | | |
|---|------------|--------------|
| ## IndustryHealth Care and Social Assistance | 1.073e+00 | 3.121e-02 |
| ## IndustryInformation | 3.241e-01 | 4.571e-02 |
| ## IndustryManufacturing | 6.372e-01 | 2.698e-02 |
| ## IndustryMining | 1.928e+00 | 1.426e-01 |
| ## IndustryOther Services (except Public Administration) | 2.230e-01 | 2.634e-02 |
| ## IndustryProfessional, Scientific, and Technical Services | 6.607e-01 | 2.650e-02 |
| ## IndustryReal Estate Rental and Leasing | -1.802e-01 | 4.531e-02 |
| ## IndustryRetail Trade | 1.825e-01 | 2.308e-02 |
| ## IndustryTransportation and Warehousing | 6.206e-01 | 3.255e-02 |
| ## IndustryWholesale Trade | 5.501e-01 | 2.871e-02 |
| ## recession_20071 | 4.204e+00 | 5.994e-01 |
| ## RegionNE:UrbanRural2 | 6.429e-02 | 4.196e-02 |
| ## RegionSouth:UrbanRural2 | 8.270e-02 | 3.929e-02 |
| ## RegionWest:UrbanRural2 | 7.780e-02 | 4.331e-02 |
| ## p_GrAppv:I(Term <= 2)TRUE | -1.278e+00 | 9.616e-02 |
| ## p_GrAppv:I(Term == 5)TRUE | -4.129e-01 | 1.042e-01 |
| ## p_GrAppv:I(Term == 7)TRUE | -7.563e-02 | 1.431e-01 |
| ## p_GrAppv:I(Term == 10)TRUE | -3.074e+00 | 6.338e-01 |
| ## p_GrAppv:Term | 1.667e-02 | 4.741e-02 |
| ## p_GrAppv:I(Term^2) | 5.857e-03 | 4.374e-03 |
| ## I(Term <= 2)TRUE:recession_20071 | -1.070e+00 | 4.037e-01 |
| ## I(Term == 5)TRUE:recession_20071 | -5.269e-01 | 2.939e-01 |
| ## I(Term == 7)TRUE:recession_20071 | -2.694e-01 | 4.128e-01 |
| ## Term:recession_20071 | -9.935e-01 | 2.401e-01 |
| ## I(Term^2):recession_20071 | 3.563e-02 | 2.366e-02 |
| ## p_GrAppv:recession_20071 | -1.447e+00 | 3.220e-01 |
| ## RegionNE:recession_20071 | -8.640e-01 | 3.719e-02 |
| ## RegionSouth:recession_20071 | -3.503e-01 | 3.565e-02 |
| ## RegionWest:recession_20071 | -1.004e+00 | 3.636e-02 |
| ## p_GrAppv:I(Term <= 2)TRUE:recession_20071 | 5.648e-01 | 2.152e-01 |
| ## p_GrAppv:I(Term == 5)TRUE:recession_20071 | 3.219e-01 | 1.632e-01 |
| ## p_GrAppv:I(Term == 7)TRUE:recession_20071 | 4.220e-01 | 2.401e-01 |
| ## p_GrAppv:Term:recession_20071 | 2.258e-01 | 1.312e-01 |
| ## p_GrAppv:I(Term^2):recession_20071 | -3.430e-03 | 1.321e-02 |
| ## | | |
| ## (Intercept) | 7.921 | 2.37e-15 *** |
| ## RegionNE | 30.261 | < 2e-16 *** |
| ## RegionSouth | -13.124 | < 2e-16 *** |
| ## RegionWest | -2.399 | 0.01645 * |
| ## UrbanRural2 | 16.254 | < 2e-16 *** |
| ## p_GrAppv | -14.999 | < 2e-16 *** |
| ## I(p_GrAppv^2) | 15.319 | < 2e-16 *** |
| ## I(p_GrAppv^3) | -13.631 | < 2e-16 *** |
| ## RevLineCrN | -27.461 | < 2e-16 *** |
| ## RevLineCrY | -7.563 | 3.93e-14 *** |
| ## LowDocY | 16.183 | < 2e-16 *** |
| ## I(Term <= 2)TRUE | 19.550 | < 2e-16 *** |
| ## I(Term == 5)TRUE | 23.144 | < 2e-16 *** |
| ## I(Term == 7)TRUE | 20.167 | < 2e-16 *** |
| ## I(Term == 10)TRUE | 9.749 | < 2e-16 *** |
| ## Term | 6.316 | 2.68e-10 *** |
| ## I(Term^2) | -2.415 | 0.01573 * |
| ## log_noEmp | 43.604 | < 2e-16 *** |
| ## NewExist2 | 15.482 | < 2e-16 *** |
| ## IndustryAdministrative and Remediation Services | 15.248 | < 2e-16 *** |
| ## IndustryAgriculture, Forestry, Fishing and Hunting | 13.527 | < 2e-16 *** |
| ## IndustryArts, Entertainment, and Recreation | 3.912 | 9.15e-05 *** |
| ## IndustryConstruction | 18.527 | < 2e-16 *** |
| ## IndustryEducational Services | 5.669 | 1.43e-08 *** |
| ## IndustryFinance and Insurance | 0.012 | 0.99077 |
| ## IndustryHealth Care and Social Assistance | 34.373 | < 2e-16 *** |
| ## IndustryInformation | 7.090 | 1.34e-12 *** |
| ## IndustryManufacturing | 23.620 | < 2e-16 *** |
| ## IndustryMining | 13.522 | < 2e-16 *** |
| ## IndustryOther Services (except Public Administration) | 8.465 | < 2e-16 *** |
| ## IndustryProfessional, Scientific, and Technical Services | 24.930 | < 2e-16 *** |
| ## IndustryReal Estate Rental and Leasing | -3.977 | 6.99e-05 *** |
| ## IndustryRetail Trade | 7.908 | 2.63e-15 *** |

```

## IndustryTransportation and Warehousing          19.064 < 2e-16 ***
## IndustryWholesale Trade                      19.160 < 2e-16 ***
## recession_20071                               7.014 2.32e-12 ***
## RegionNE:UrbanRural2                         1.532 0.12550
## RegionSouth:UrbanRural2                     2.105 0.03530 *
## RegionWest:UrbanRural2                       1.796 0.07248 .
## p_GrAppv:I(Term <= 2)TRUE                  -13.290 < 2e-16 ***
## p_GrAppv:I(Term == 5)TRUE                   -3.961 7.45e-05 ***
## p_GrAppv:I(Term == 7)TRUE                   -0.528 0.59722
## p_GrAppv:I(Term == 10)TRUE                  -4.850 1.23e-06 ***
## p_GrAppv:Term                                0.352 0.72520
## p_GrAppv:I(Term^2)                          1.339 0.18061
## I(Term <= 2)TRUE:recession_20071           -2.650 0.00805 **
## I(Term == 5)TRUE:recession_20071            -1.793 0.07298 .
## I(Term == 7)TRUE:recession_20071            -0.653 0.51400
## Term:recession_20071                         -4.138 3.50e-05 ***
## I(Term^2):recession_20071                   1.506 0.13206
## p_GrAppv:recession_20071                   -4.493 7.03e-06 ***
## RegionNE:recession_20071                   -23.231 < 2e-16 ***
## RegionSouth:recession_20071                -9.826 < 2e-16 ***
## RegionWest:recession_20071                 -27.622 < 2e-16 ***
## p_GrAppv:I(Term <= 2)TRUE:recession_20071   2.625 0.00867 **
## p_GrAppv:I(Term == 5)TRUE:recession_20071   1.973 0.04854 *
## p_GrAppv:I(Term == 7)TRUE:recession_20071   1.758 0.07878 .
## p_GrAppv:Term:recession_20071              1.721 0.08519 .
## p_GrAppv:I(Term^2):recession_20071         -0.260 0.79515
## ---
## 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: 188157 on 307303 degrees of freedom
## AIC: 188275
##
## 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

```

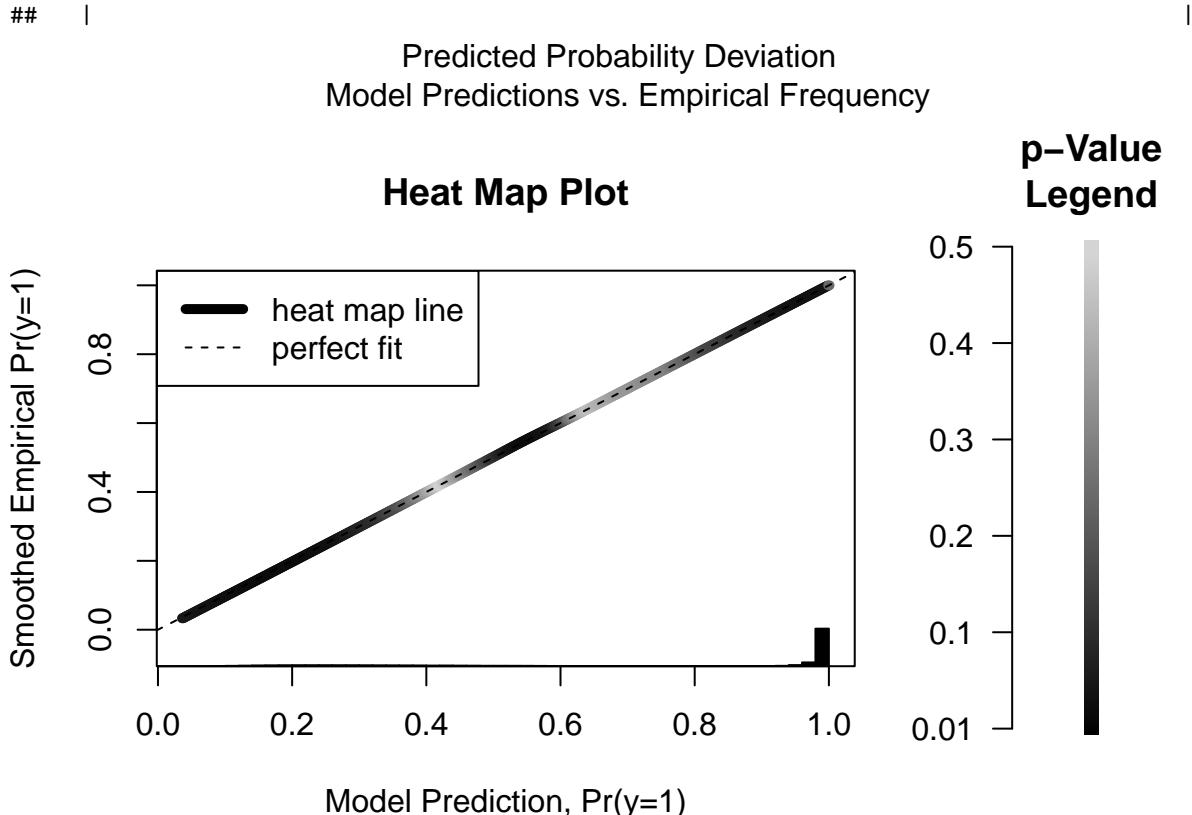
The heatmap diagnostic in [Esarey and Pierce \(2012\)](#) suggests that our final model fits the data pretty well. The tests suggest a statistical lack of fit, but the heatmap plot does reveal any practical lack of fit. Model predictions is closely aligned with empirical success probabilities.

```

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 0.8517 is much higher than the no information rate of 0.7081. Importantly, the [sensitivity](#) and [specificity](#) are, respectively, 0.8047 and 0.8711.

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

## Confusion Matrix and Statistics
##
```

| | | Reference | |
|------------|---|-----------|-------|
| | | 0 | 1 |
| Prediction | 0 | 18048 | 7016 |
| | 1 | 4379 | 47398 |

```
##
```

Accuracy : 0.8517
95% CI : (0.8492, 0.8542)

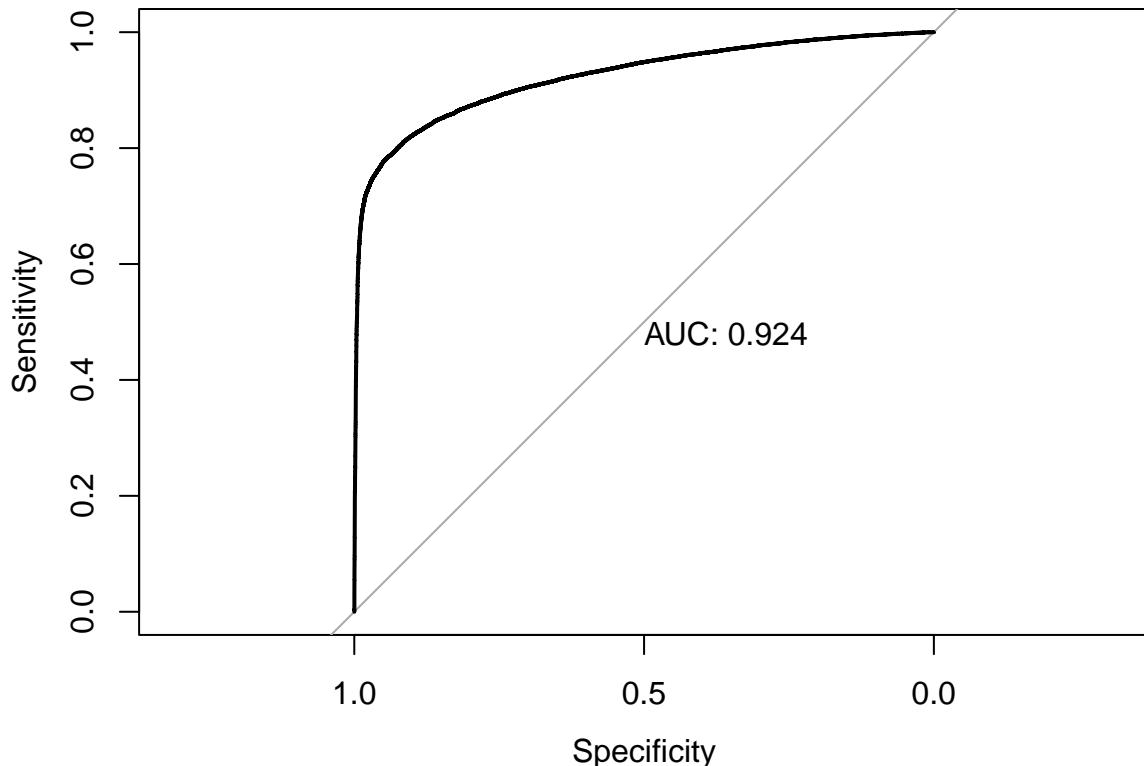
```

##      No Information Rate : 0.7081
##      P-Value [Acc > NIR] : < 2.2e-16
##
##              Kappa : 0.6532
##
## McNemar's Test P-Value : < 2.2e-16
##
##      Sensitivity : 0.8047
##      Specificity : 0.8711
##      Pos Pred Value : 0.7201
##      Neg Pred Value : 0.9154
##      Prevalence : 0.2919
##      Detection Rate : 0.2349
##      Detection Prevalence : 0.3262
##      Balanced Accuracy : 0.8379
##
##      '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 lower classification accuracy (0.8087) than our logistic regression model (0.8517). And 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
##

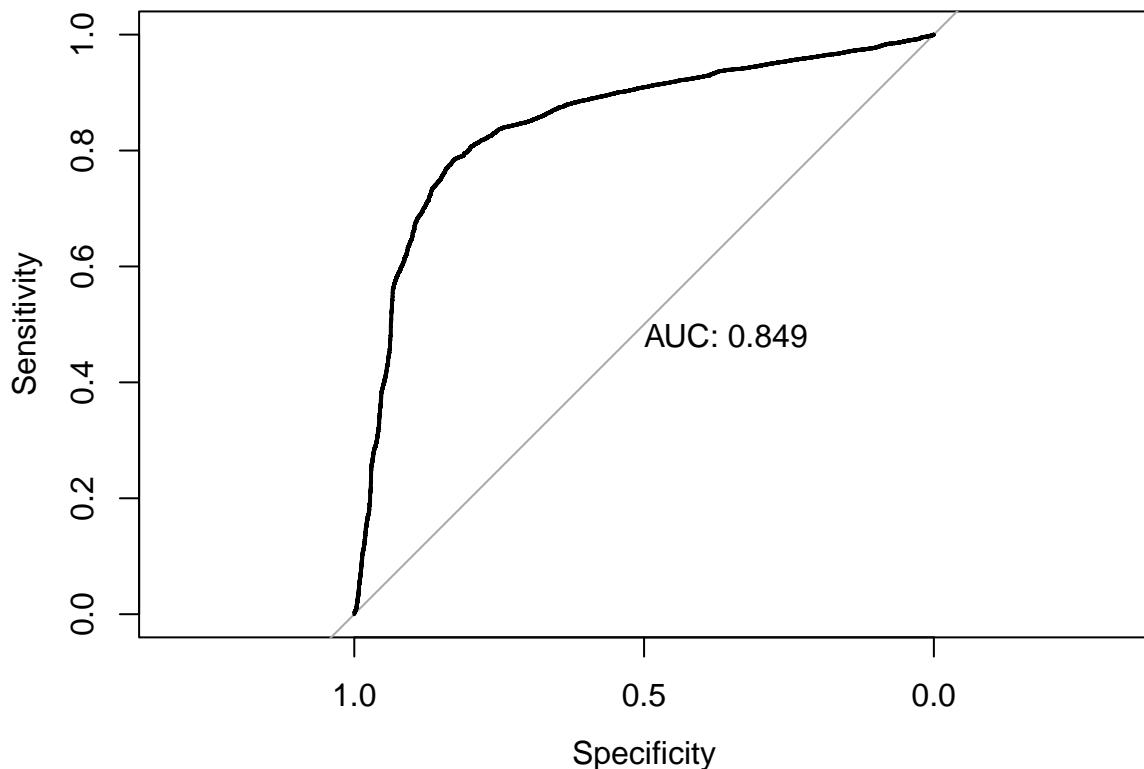
```

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) over our logistic regression model (0.8517).

```

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
##

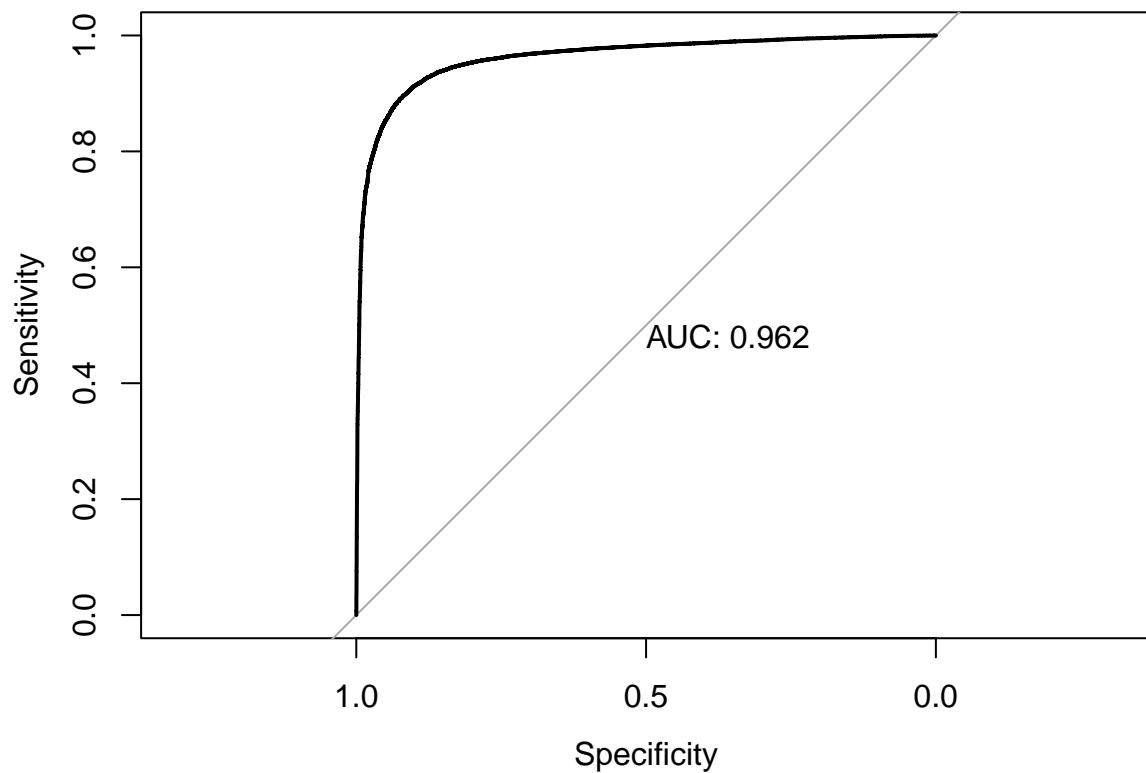
```

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)

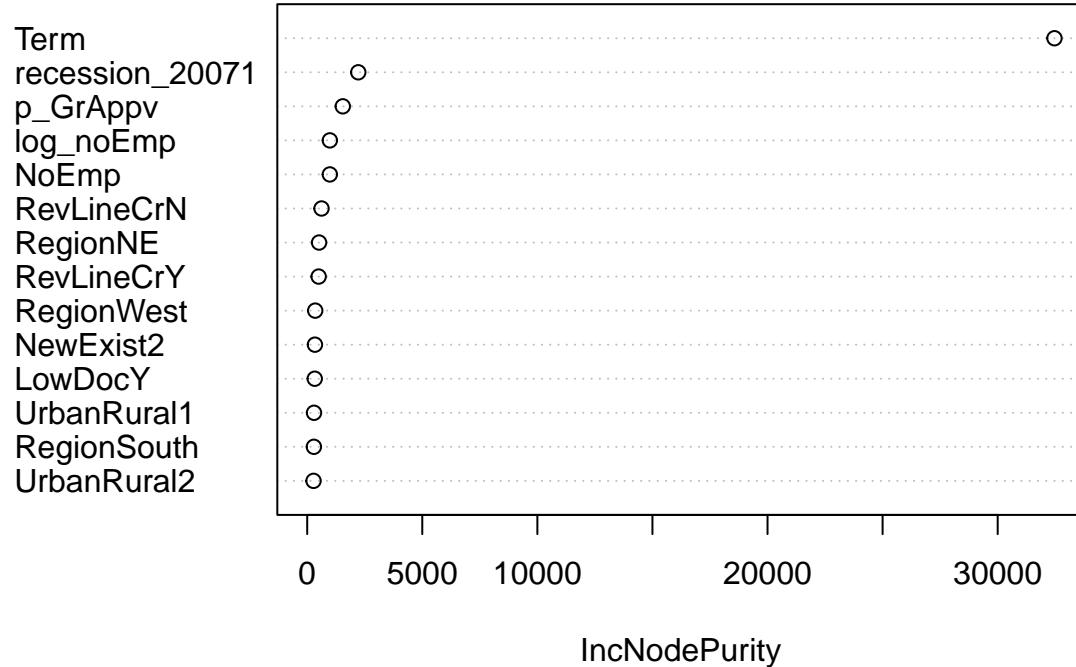
```



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

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

rforest



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 favorable for the random forest in terms of classification accuracy. Perhaps our logistic regression model can be improved by considering more variables, higher order interaction terms, and/or additional data splits. In the Appendix we go a different direction and explore an ensemble inspired 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.
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)

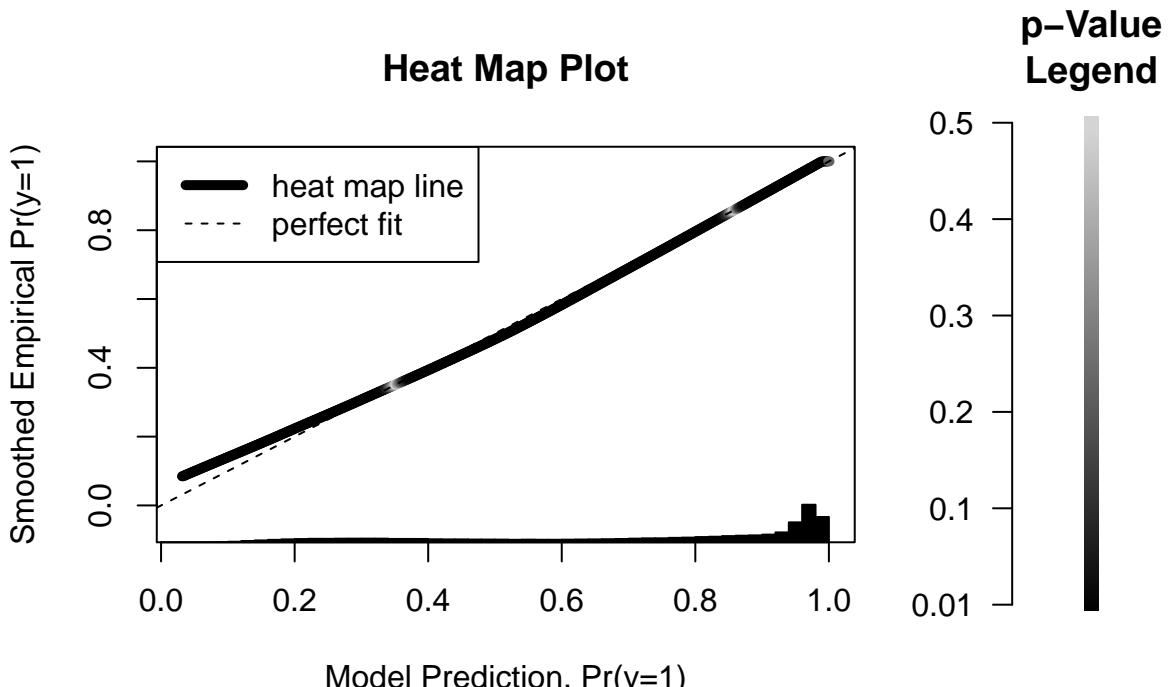
Here we refit our logistic regression model without the discrete Term effects. This model exhibits severe lack of fit.

```
m2 <- glm(MIS_Status ~ Region*UrbanRural +
  p_GrAppv + I(p_GrAppv^2) + I(p_GrAppv^3) +
  RevLineCr + LowDoc + (Term + I(Term^2))*p_GrAppv + log_noEmp +
  NewExist + Industry +
  recession_2007*((Term + I(Term^2))*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 1961 weighted bins.
##
## Calculating optimal loess bandwidth...
## aicc Chosen Span = 0.8981399
##
## Generating Bootstrap Predictions...
## |
```

Predicted Probability Deviation
Model Predictions vs. Empirical Frequency



```
##
##
## ****
## 95.08365% of Observations have one-tailed p-value <= 0.1
## Expected Maximum = 20%
## ****
```

Appendix (ensemble model inspired 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))*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
##     8.732   0.265   8.999
```

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 19416  3301
##           1  3011 51113
##
##          Accuracy : 0.9179
##             95% CI : (0.9159, 0.9198)
##    No Information Rate : 0.7081
##    P-Value [Acc > NIR] : < 2.2e-16
##
##          Kappa : 0.802
##
##  Mcnemar's Test P-Value : 0.0002752
##
##          Sensitivity : 0.8657
##          Specificity : 0.9393
##    Pos Pred Value : 0.8547
##    Neg Pred Value : 0.9444
##          Prevalence : 0.2919
##    Detection Rate : 0.2527
```

```
##      Detection Prevalence : 0.2956
##      Balanced Accuracy : 0.9025
##
##      'Positive' Class : 0
##
sba_roc_rf = roc(y_test ~ p_super_test,
                  plot = TRUE, print.auc = TRUE)
```

