SMEs
Loan Risk Detection

Exploratory Data Analysis & Predictive Model

Witchayut Chuaychukul	**63070501057**
Suppakorn Rakna	**63070501061**
Siriwat Chotilersak	**63070501073**
Aussadawut Ardrit	**63070501084**

Introduction & Analytic objective

Small and Medium Enterprises (SMEs)

Carry out activities based on the entrepreneur's funds

Independent

Low-investment

Small employee number

High chance of being affected by volatile economic conditions and business disruptions

Getting a loan from a reliable organization

There are SMEs who are unable to pay off their debts from applying for loans or Securities as collateral as specified

Find out how financial aid organizations should provide funding to SMEs or not? **By detect risk**

Data description and preparation.

Data description

	Promoting and assisting small enterprises in the U.S. credit market	Insurance providers reduce the risk for a bank by taking on some of the risks by guaranteeing a portion of the loan	There have been many success stories of start-ups receiving SBA loan guarantees such as FedEx and Apple Computer

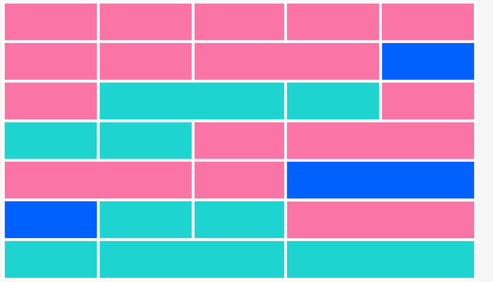
U.S. Small Business Administration

1987 - 2014

899,164 records

LoanNr_ChkDgt Primary key	Name Borrower	City	State	Zip
Bank	BankState	NAICS North American industry classification system code		ApprovalDate
ApprovalFY Approval fiscal year	Term Loan contract period (#month)		NoEmp Amount of employee	NewExist 1 = Existing, 2 = New
CreateJob	RetainedJob	FranchiseCode 0/1 = no franchise	UrbanRural 1 = Urban, 2 = rural, 0 = undefined	
RevLineCr Revolving line of credit: Y = Yes, N = No		LowDoc LowDoc Loan Program	ChgOffDate The date when a loan is declared to be in default	
DisbursementDate	DisbursementGross Amount disbursed	BalanceGross Gross amount outstanding	MIS_Status Loan status charged off = CHGOFF, Paid in full = PIF	
ChgOffPrinGr Charged-off amount	GrAppv Gross amount of loan approved by bank		SBA_Appv SBA's guaranteed amount of approved loan	

Data preparation Functions



Determine if date is in between the economic recession

```
is_between <- function(date){
  return(between(
    date,
    as.Date("2007-12-01"),
    as.Date("2009-06-30")
  ))
}
```

DisbursementDate

Get value from currency $xx,xxx

```
get_value <- function(str) {
  val <- as.character(str) %>%
    str_replace_all("$", "") %>%
    str_replace_all(",", "") %>%
    substring(2)
  return(as.double(val))
}
```

GrAppv | SBA_Appv

DisbursementGross

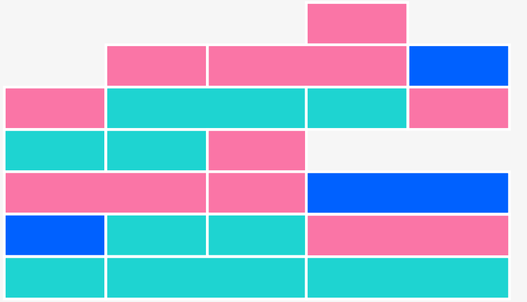
Get Date from chr ex. “30-May-22”

```
get_date <- function(str) {
  list <- strsplit(str, split = "-")[[1]]
  date <- list[1]
  month <- match(list[2],
         substr(month.name, 0, 3))
  year <- ifelse(list[3] > 50,
                 paste("19", list[3], sep = ""),
                 paste("20", list[3], sep = "")
  )
  return(as.Date(
   paste(year, month, date, sep = “-“))
   )
}
```

DisbursementDate

ApprovalDate

Data preparation

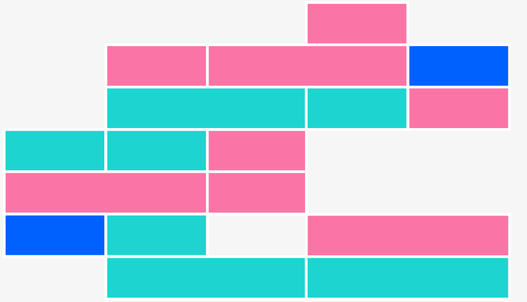


LoanNr_ChkDgt Primary key	Name Borrower	City	Zip	Bank
No potential		Use `State ` To reduce factor level		High factor level (**5803** Levels)

UrbanRural 1 = Urban, 2 = rural, 0 = undefined	High amount of UNDEFINED

	UrbanRural <fct>	n <int>
1	**0**	**109101**
2	1	285673
3	2	63055

Data preparation



~~ChgOffDate~~

~~The date when a loan is declared to be in default~~

~~ChgOffPrinGr~~

~~Charged-off amount~~

Missing values (81.91%)

```
> nrow(filter(data, ChgOffDate==''))
[1] 736465
```

~~ApprovalFY~~

~~Approval fiscal year~~

Not in the area of interest

~~BalanceGross~~

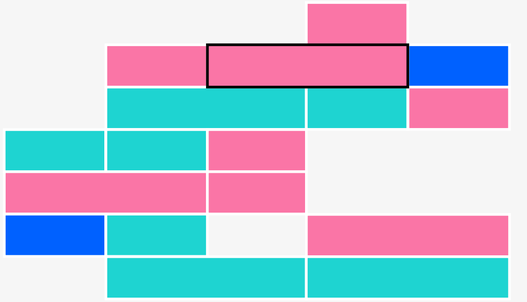
~~Gross amount outstanding~~

Very right skewed or error in data collection

```
> data %>%
+ select(BalanceGross) %>%
+ mutate(BalanceGross = get_value(BalanceGross)) %>%
+ arrange(desc(BalanceGross)) %>%
+ head(20)
 BalanceGross
1        996262
2        827875
3        395476
4        115820
5         96908
6         84617
7         43127
8         41509
9         37100
10        25000
11        12750
12         9111
13         1760
14          600
15            0
16            0
17            0
18            0
19            0
20            0
```

Then, `drop_na()` for the first time to remove other latent NA

Data preparation



Scale to **sector level**

NAICS

North American industry classification system code

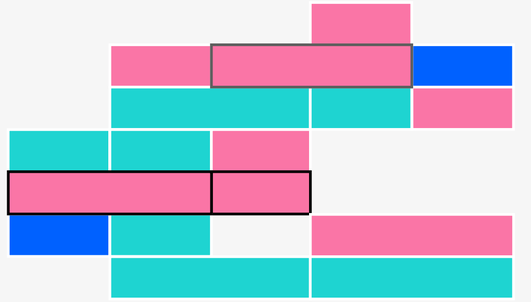
```
> head(data$NAICS)
[1] 233210 321999 621999 445310 0 722410
```


```
… %>%
mutate(NAICS = substr(NAICS, 0, 2)) %>%
filter(NAICS != 0) %>% # remove missing values
```

Sector	Definition
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

Data preparation

Remove data entry errors and missing value



RevLineCr	LowDoc
Revolving line of credit: Y = Yes, N = No	LowDoc Loan Program

```
> as.factor(data$RevLineCr)
…
Levels:  - , . ` 0 1 2 3 4 5 7 A C N Q R T Y
```

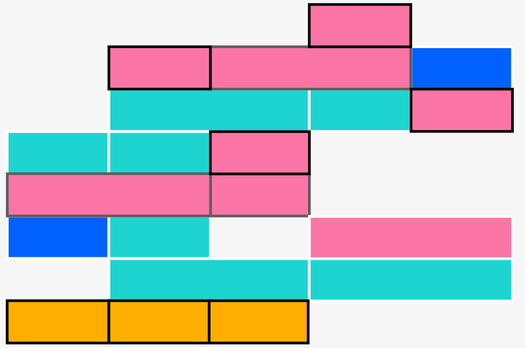
```
> as.factor(data$LowDoc)
…
Levels:  0 1 A C N R S Y
```

Keep Yes and No

```
%>% filter(… %in% c("Y", "N")) %>%
```

Data preparation

Create new variables



BankState	`# Flag field which identifies where the State is the same as the BankState`
	
	`… %>% mutate(BankInState = ifelse(State == BankState, 1, 0)) %>% …`

State	BankInState

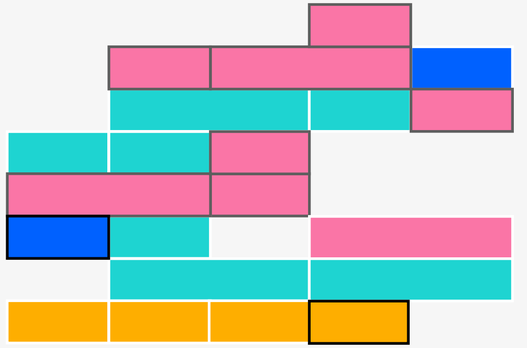
NewExist
 	
0, 1 = Existing, 2 = New
	`# New exist: less than 2 years`
	
	`… %>% mutate(NewExist = as.factor(NewExist)) %>% …`
	
	NewExist
	
	0 = Existing, 1 = New

FranchiseCode
 	
0/1 = no franchise
	`# Flag field which identifies where they have franchise or note`
	`… %>%`
	`mutate(Franchise = ifelse(`
	`  FranchiseCode %in% c("0", "1"), 0, 1))%>% …`
	
	Franchise
	
	0 = No, 1 = Have

Data preparation

Economic recession

2007-12-01 → 2009-06-30



The loans that were coded as “Recession=1” include those that were active for at least a month during the Great Recession time frame. This was calculated by adding the length of the loan term in days to the disbursement date of the loan.

`DisbursementDate`

```
… %>%
  mutate(xx = DisbursementDate + (Term * 30)) %>%
  mutate(DisbursementGross = get_value(DisbursementGross)) %>%
  mutate(Recession = ifelse(is_between(xx), 1, 0)) %>% …
```

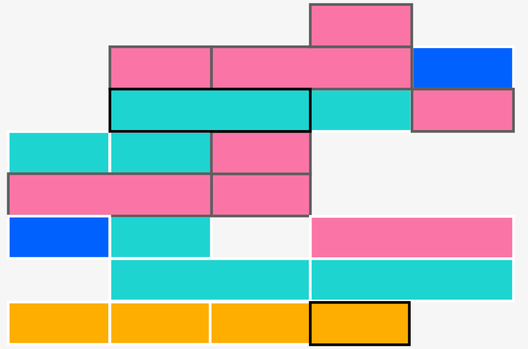
`Recession`

0 = not during, 1 = during

https://money.cnn.com/2010/09/20/news/economy/recession_over/

Data preparation

Real estate



Since the term of the loan is a function of the expected lifetime of the assets, loans backed by real estate will have terms 20 years or greater (≥240 months) and are the only loans granted for such a long term, whereas loans not backed by real estate will have terms less than 20 years (<240 months)

`Term`

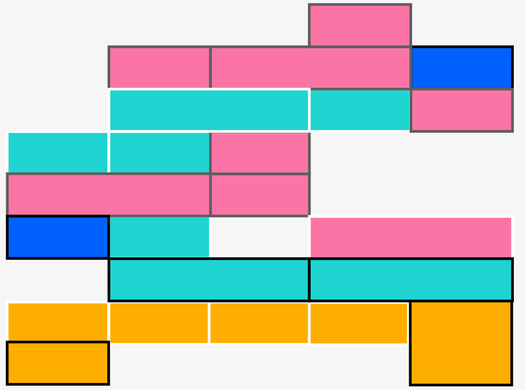
Loan contract period (#month)

```
… %>%
mutate(RealEstate = ifelse(Term >= 240, 1, 0))
%>% …
```

`RealEstate`

`0` =no, 1 = yes

Data preparation



GrAppv	SBA_Appv	Portion
Gross amount of loan approved by bank	SBA's guaranteed amount of approved loan	

```
… %>% mutate(Portion = SBA_Appv / GrAppv) %>% …
```

Hypothesis: The timing at which the funds were received could have a <u>negative relationship</u> with a business's ability to pay off

ApprovalDate
DisbursementDate

```
… %>%
mutate(DaysToDisbursement = difftime(ApprovalDate,
DisbursementDate, units = "days")) %>%
filter(DaysToDisbursement >= 0)
%>% …
```

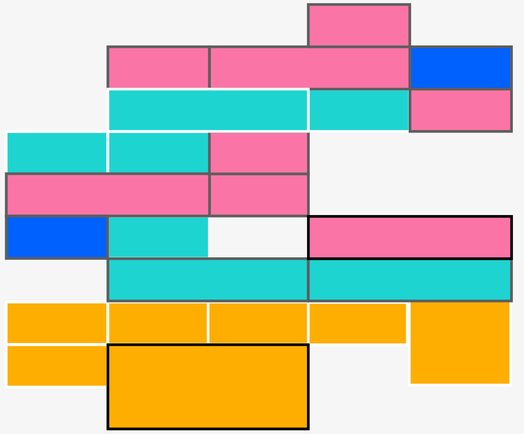
DaysToDisbursement
The number of days passed between DisbursementDate and ApprovalDate

Data preparation

Default

“no” = This loan is not risk

“yes” = This loan is risk



Find out how financial aid organizations should provide funding to SMEs or not? **By detect risk**

`MIS_Status`

Loan status charged off = CHGOFF, Paid in full = PIF

```
… %>%
  mutate(Default = case_when(
    MIS_Status == as.character("P I F") ~ 'no',
    MIS_Status == as.character("CHGOFF") ~ 'yes'
  )) %>% …
```

`Default`

New data

457,829 records

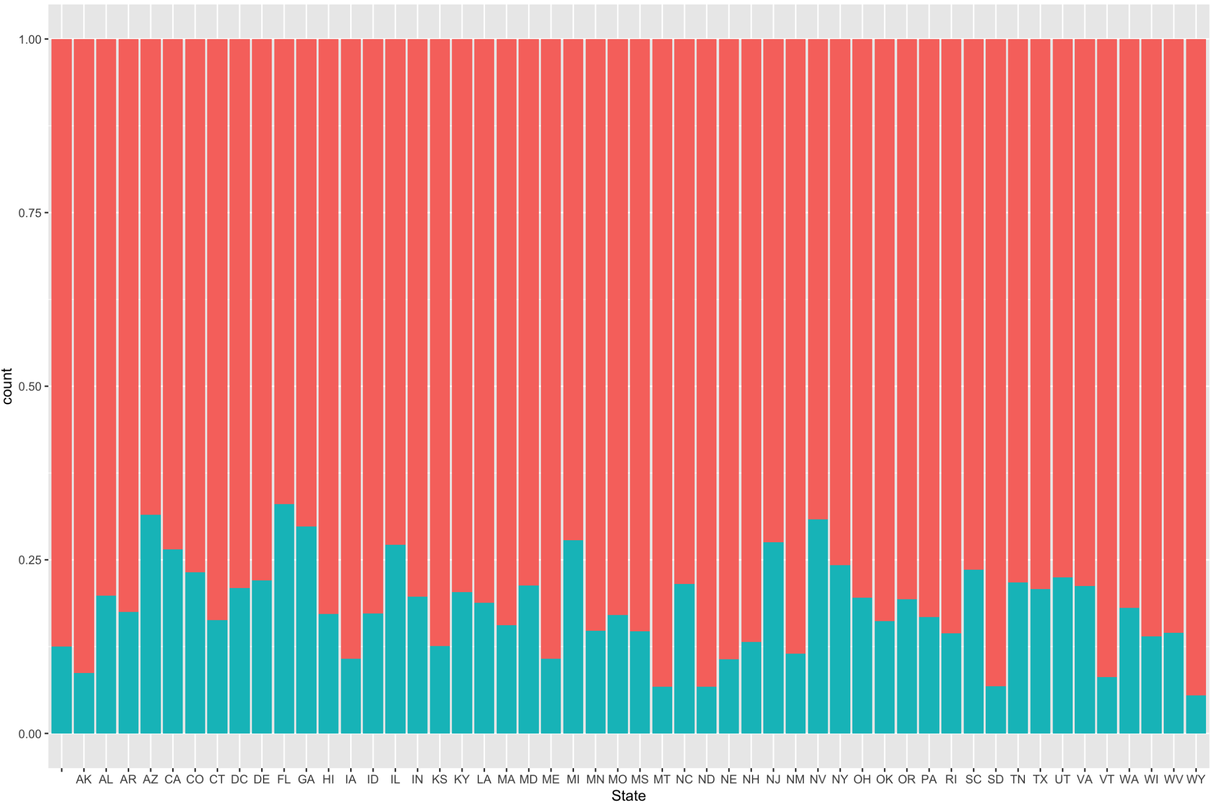
State	NAICS North American industry classification system code		Term Loan contract period (#month)	
NoEmp Amount of employee	NewExist 1 = Existing, 2 = New	CreateJob	RetainedJob	LowDoc LowDoc Loan Program
Portion	DaysToDisbursement	RevLineCr Revolving line of credit: Y = Yes, N = No		DisbursementGross Amount disbursed
BankInState	NewExist 0 = Existing, 1 = New	Franchise 0 = No, 1 = Have	Recession 0 = not during, 1 = during	RealEstate 0 =no, 1 = yes
Default				

Categorical	Numerical

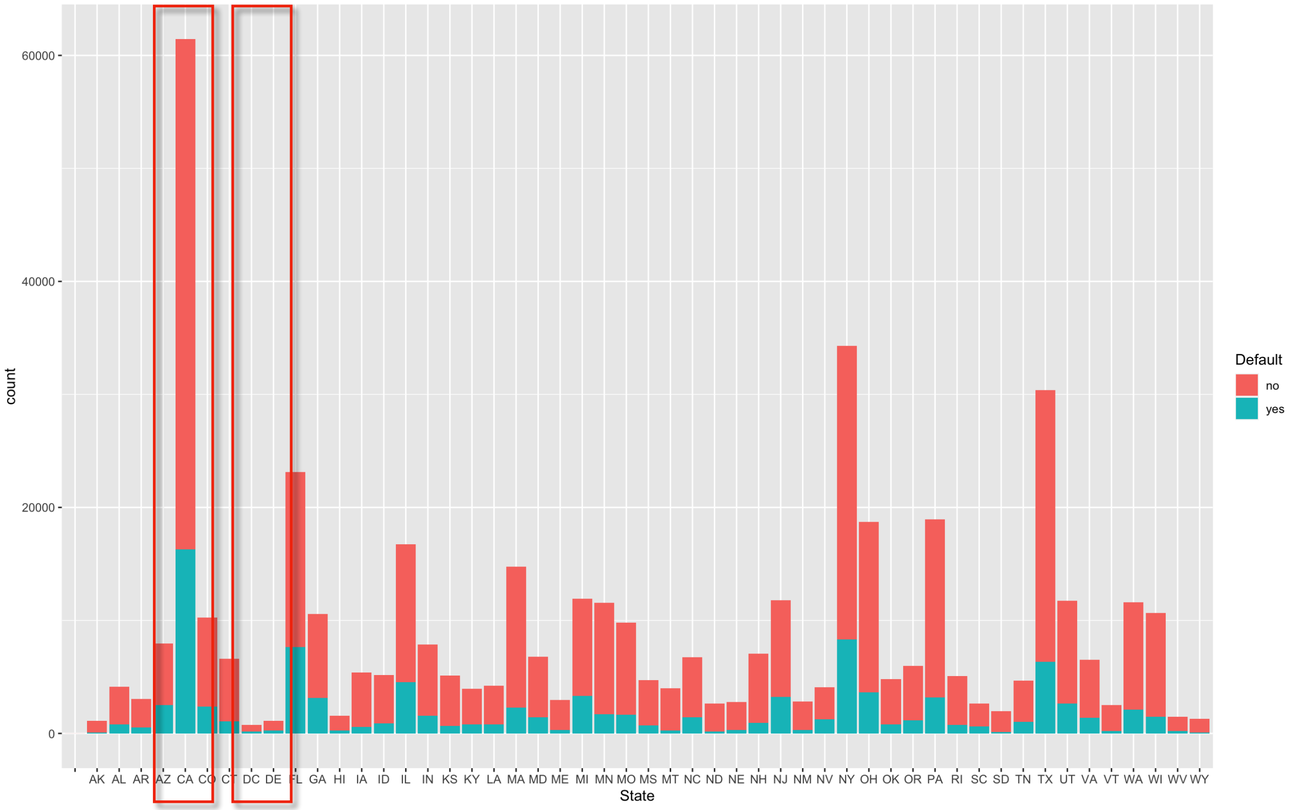
Data exploration and visualization.

State

`geom_bar(position = “fill")`



`geom_bar(position = "stack")`

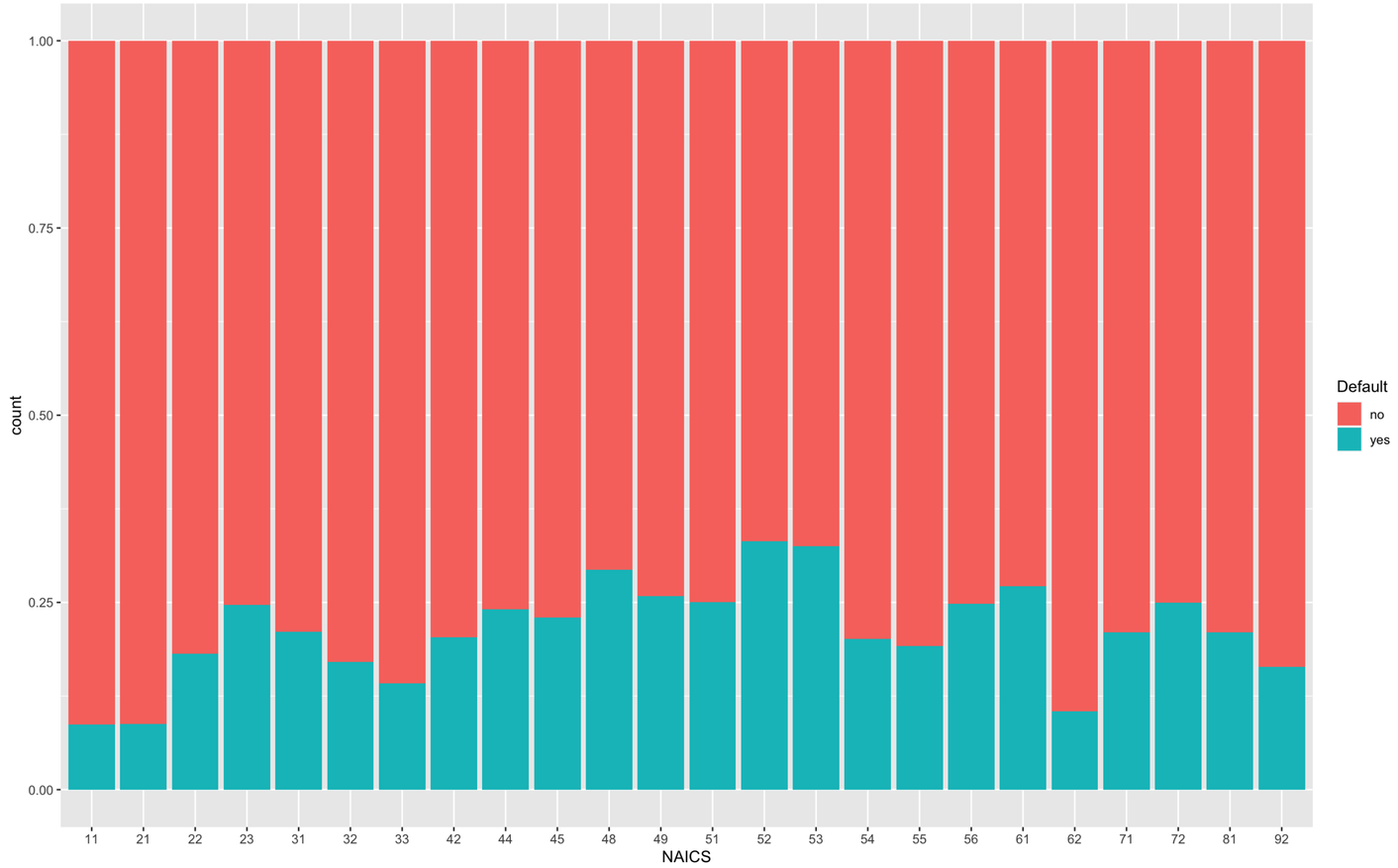


The amount of record for each state is **variant**, may cause bias in term of information.

NAICS ~ Default

North American industry classification system code

Sector	Definition
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
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

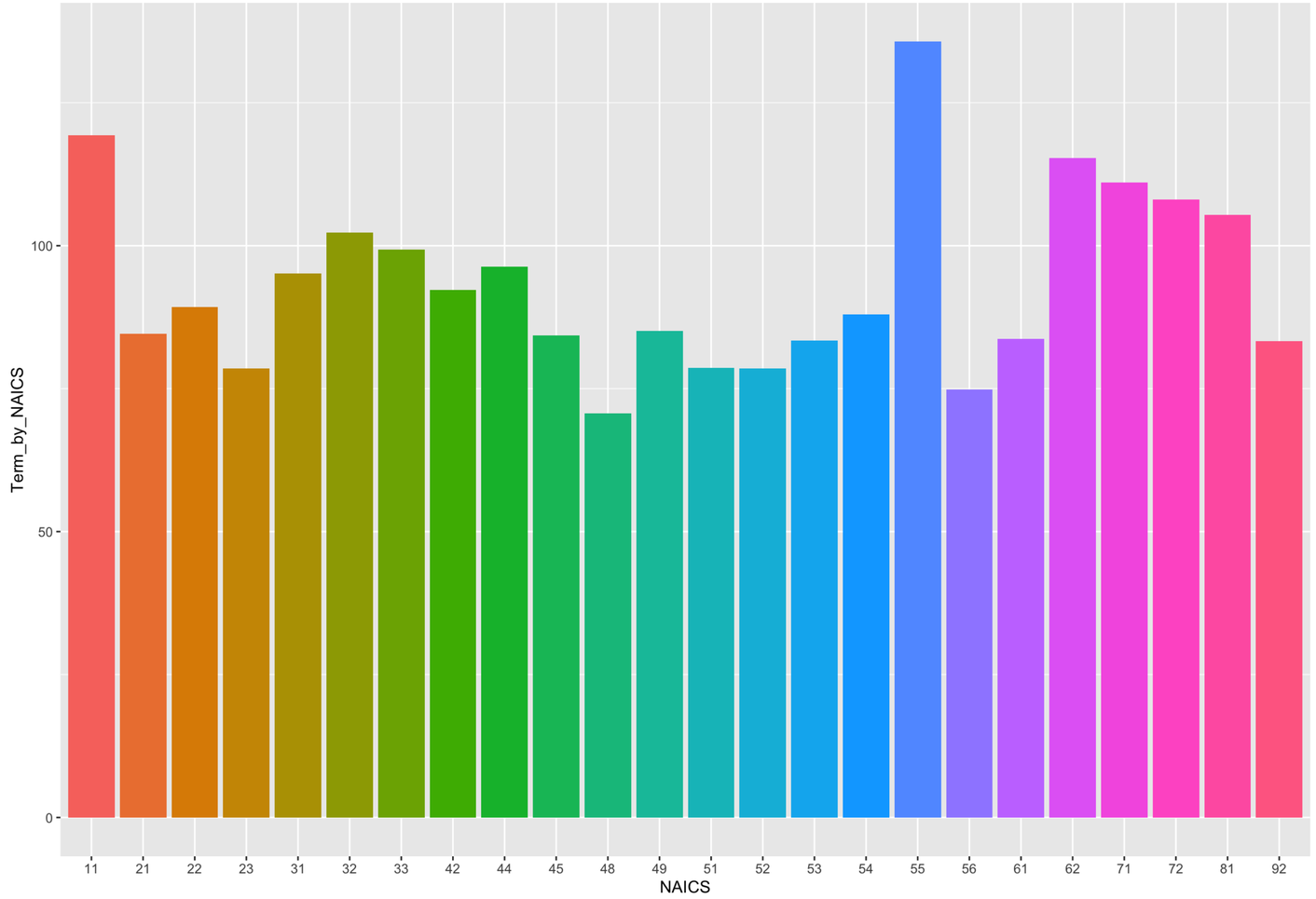


- the least risk industries such that Health Care, Mining, Agriculture, and Management.
- On the other hand, Real Estate and Finance were risk industries to approve a loan to, with about a 25% default rate.

NAICS ~ Term

North American industry classification system code

Sector	Definition
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
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

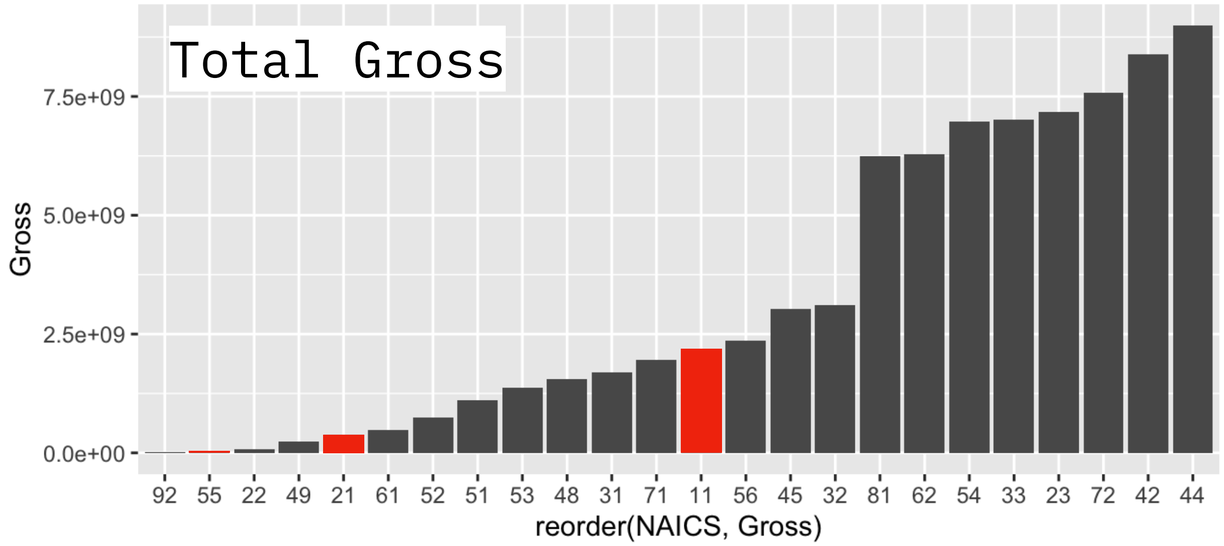


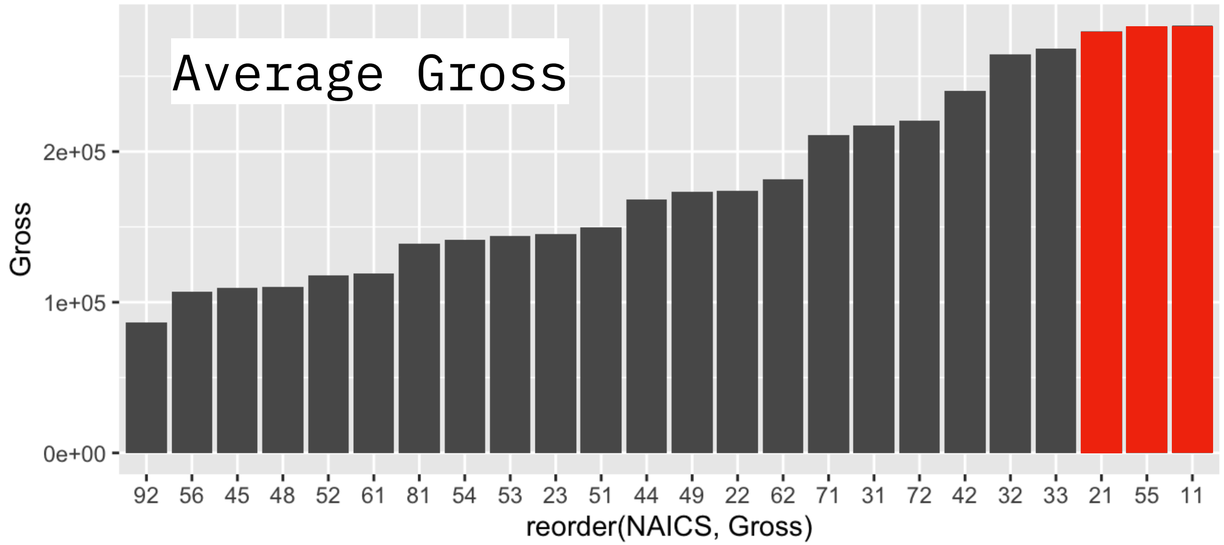
- Industries that took the longest loans were usually likely to pay.
- The Management industry took out the longest loans, followed by Health Care and Food Services. This evidence supports our assumption that loans with longer Terms are less likely to default.

NAICS ~ Gross

North American industry classification system code

Sector	Definition
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
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

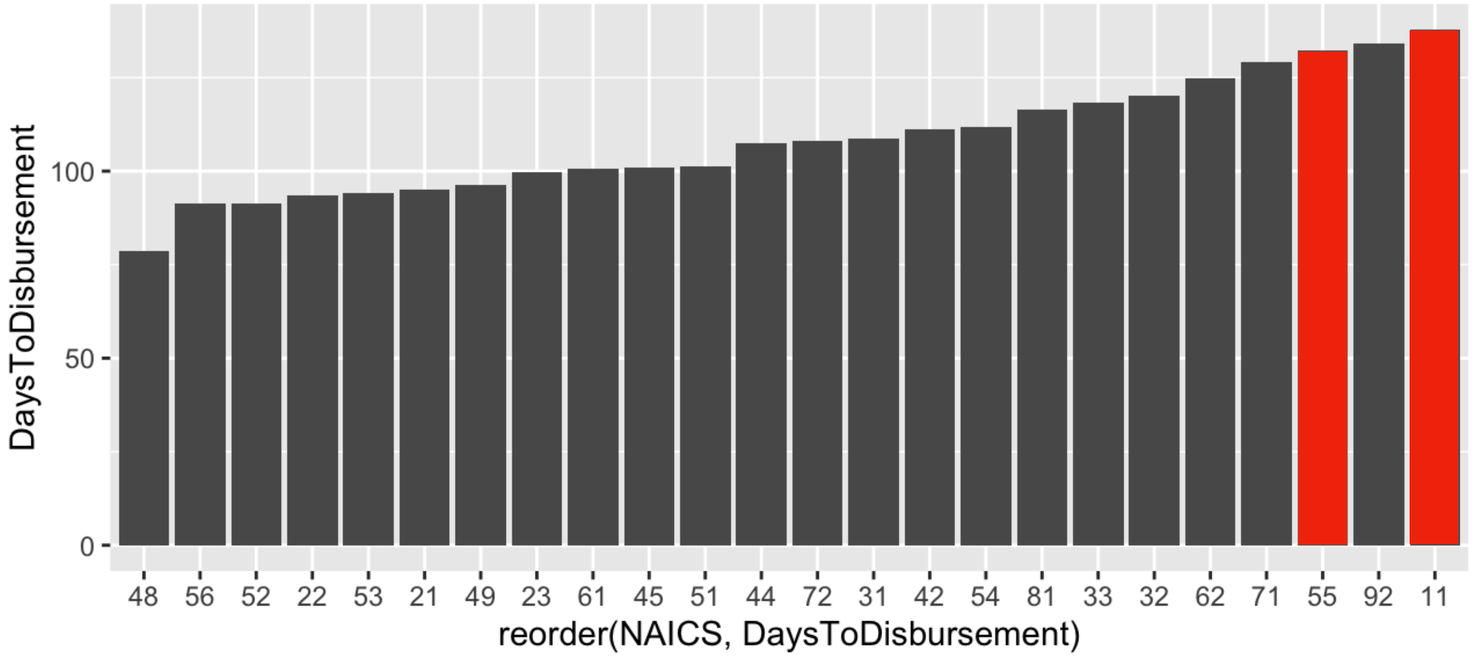




Although the Agriculture, forestry, fishing and hunting(11), Mining, quarrying, and oil and gas extraction(21), and Management(55) had a small amount of total loan, they had the highest average loan amount compared to other industries; This suggests they had a small number of large loans

DaysToDisbursement ~ NAICS

Sector	Definition
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
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

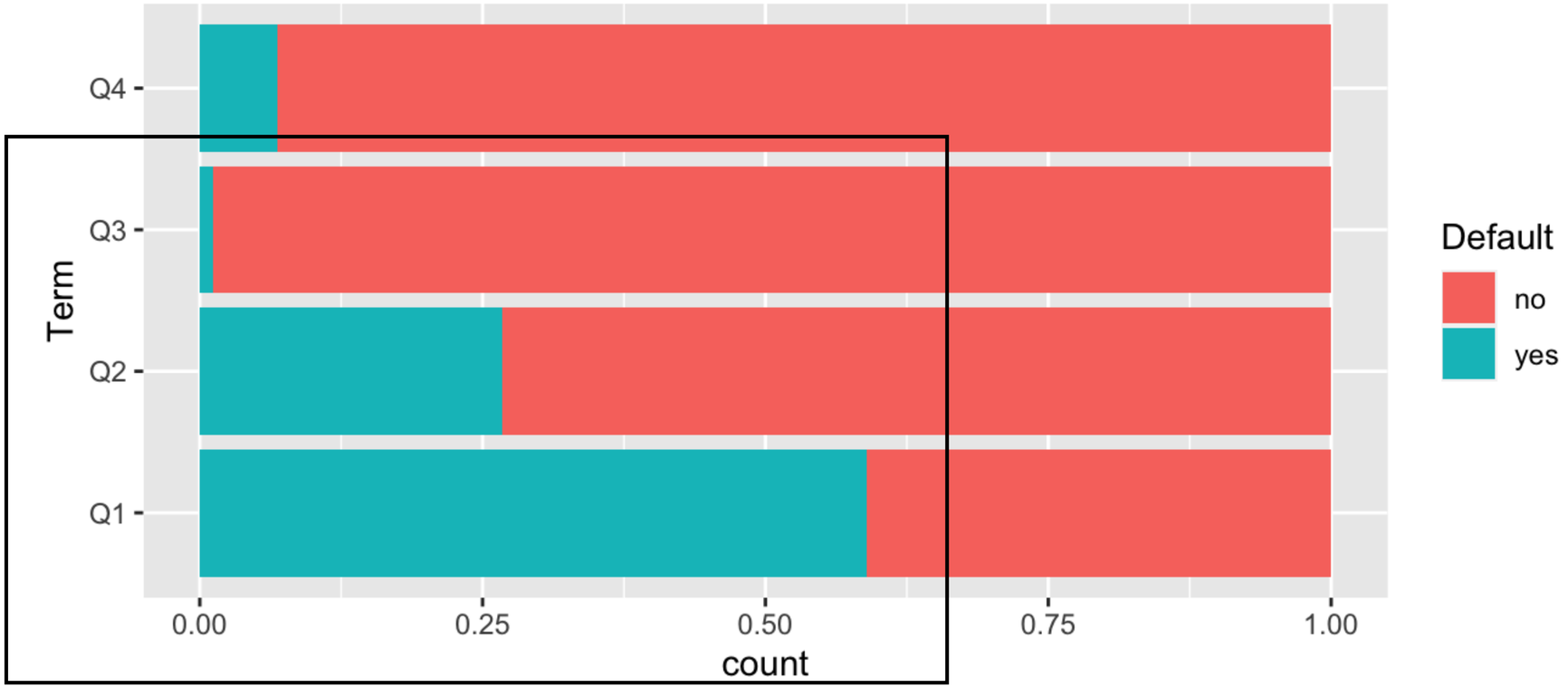


Some of the industries with the highest average loan amount also had the highest number of days to disbursement of funds

Term

Loan contract period (#month)

```
> summary(data$Term)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   59.00   84.00   93.88   90.00  527.00
```



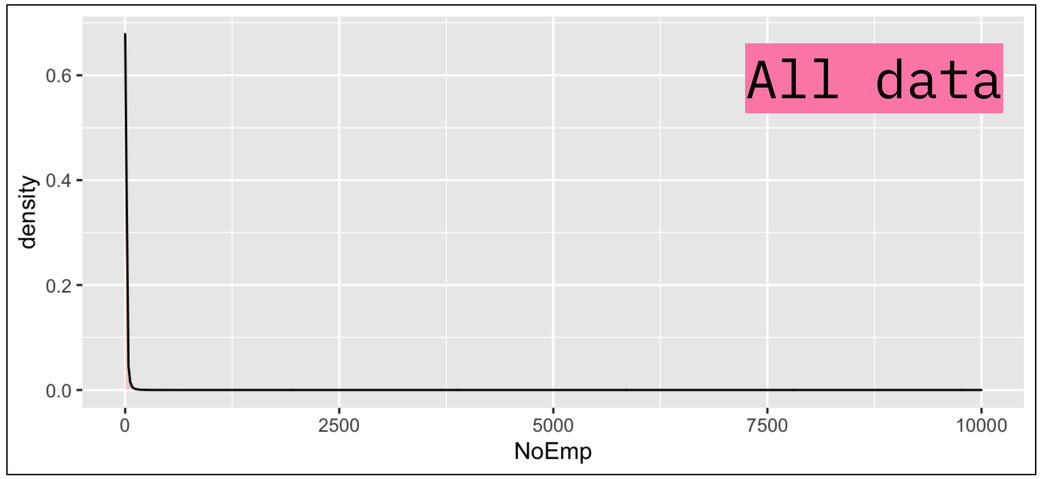
Each quantile has various ratio:

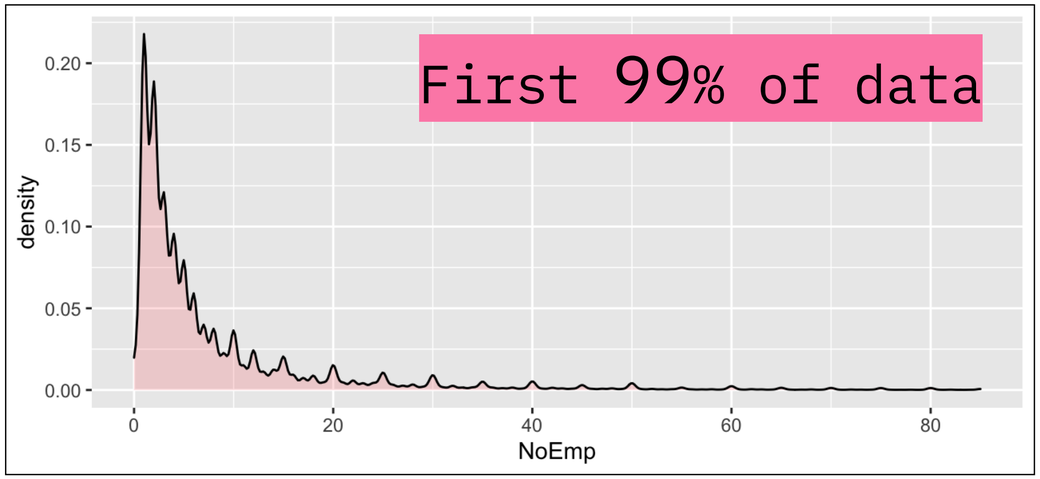
— Term has interesting outcome → recommended as predictors

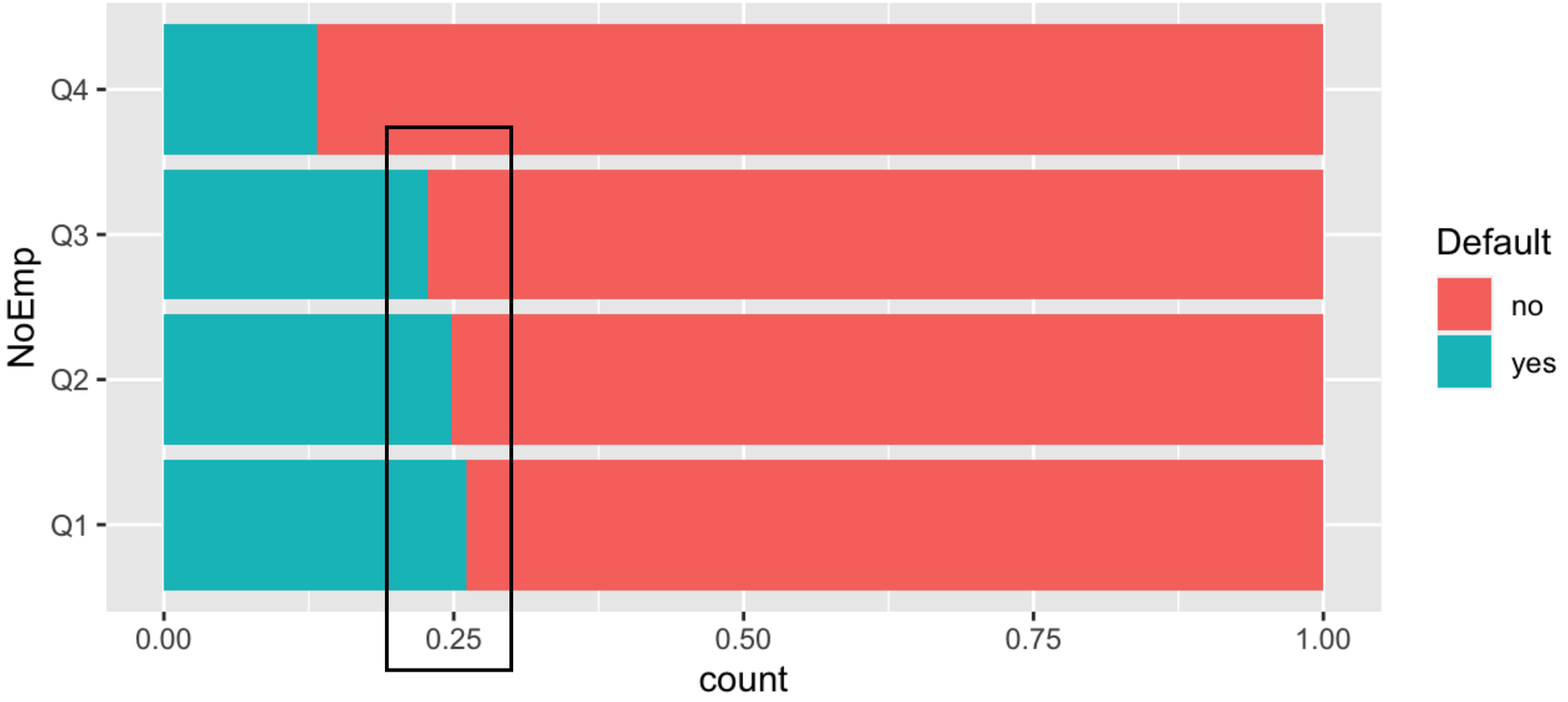
— Roughly, Term inverse to risk rate, with HIGH probability.

NoEmp

Amount of employee







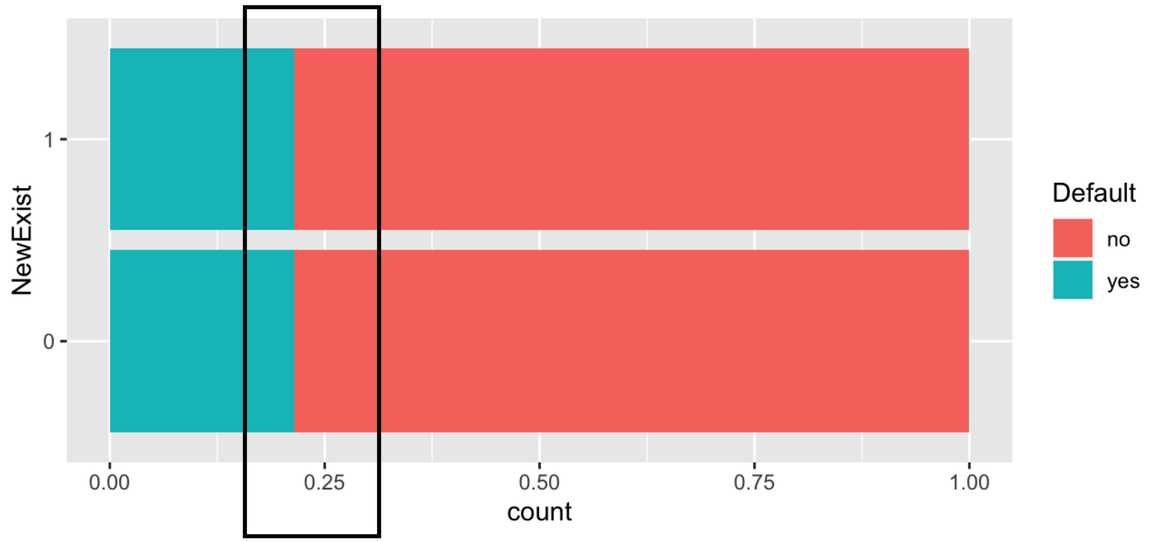
Almost no different

Risk percentage and ratio

NewExist

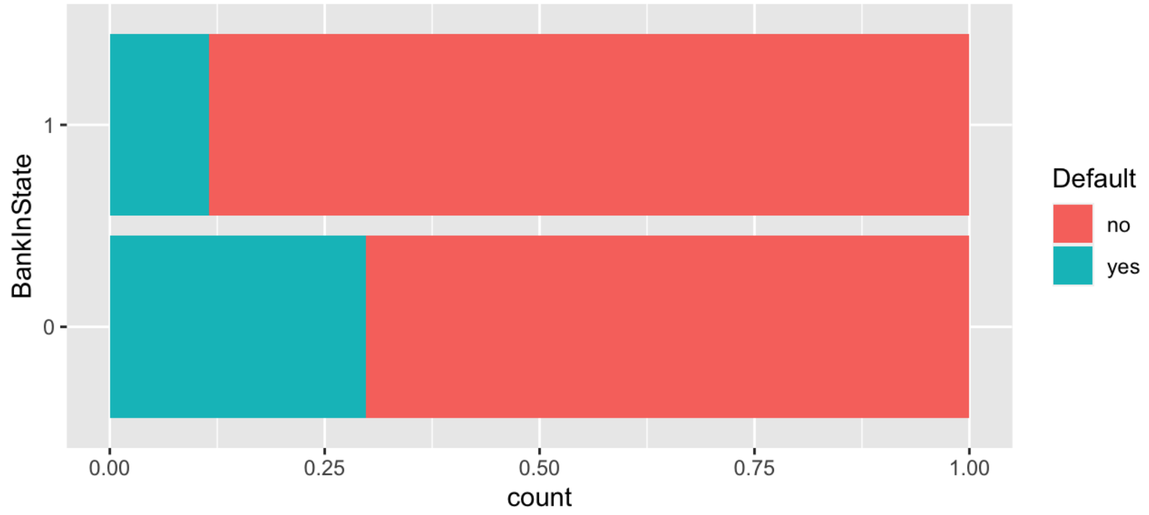
1 = Existing, 2 = New

26.46%



BankInState

45.66%

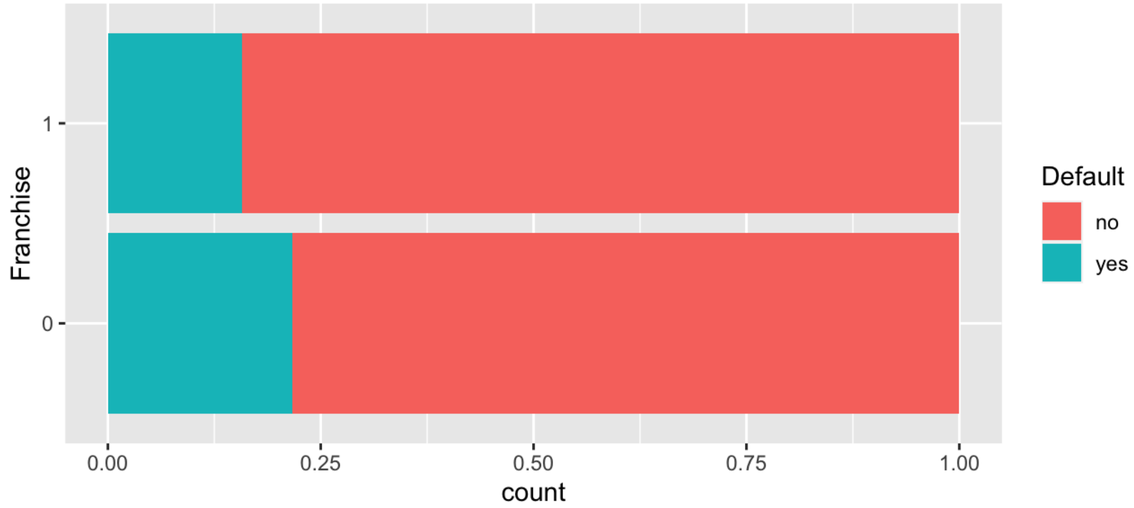


Risk percentage and ratio

Franchise

0 = No, 1 = Have

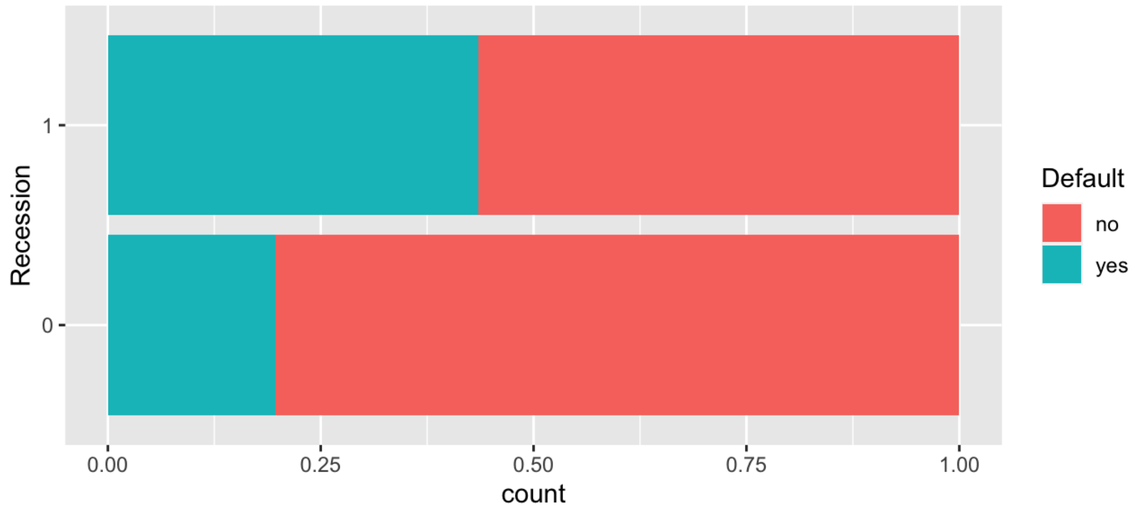
3.13%



Recession

0 = not during, 1 = during

7.70%

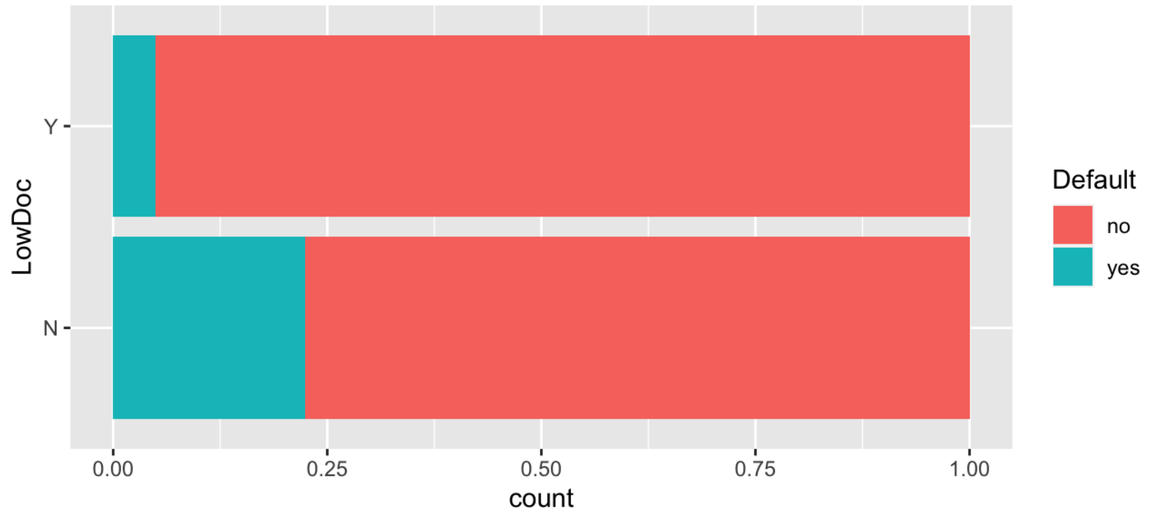


Risk percentage and ratio

LowDoc

LowDoc Loan Program

5.49%

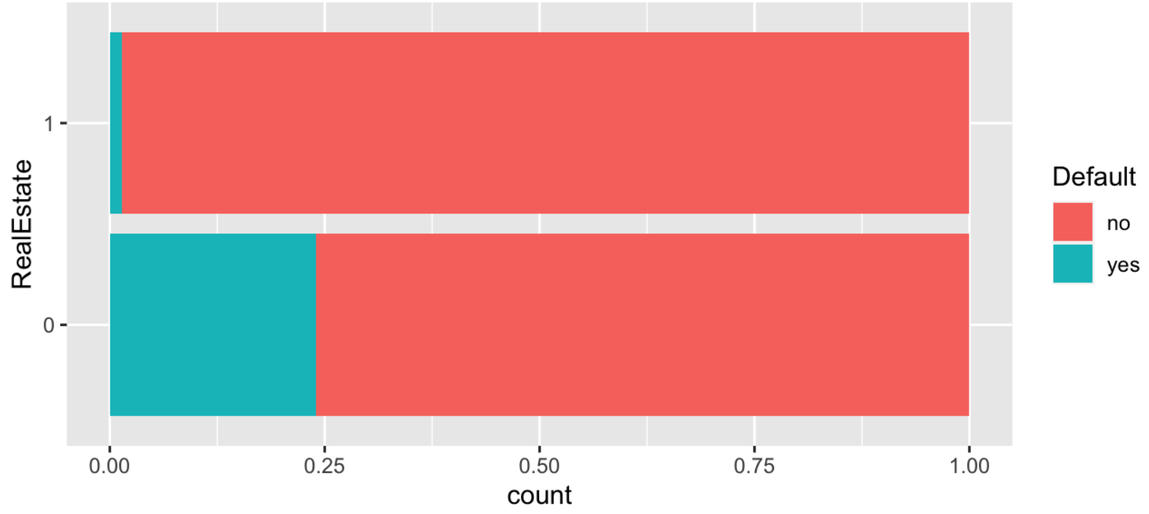


From `Term`

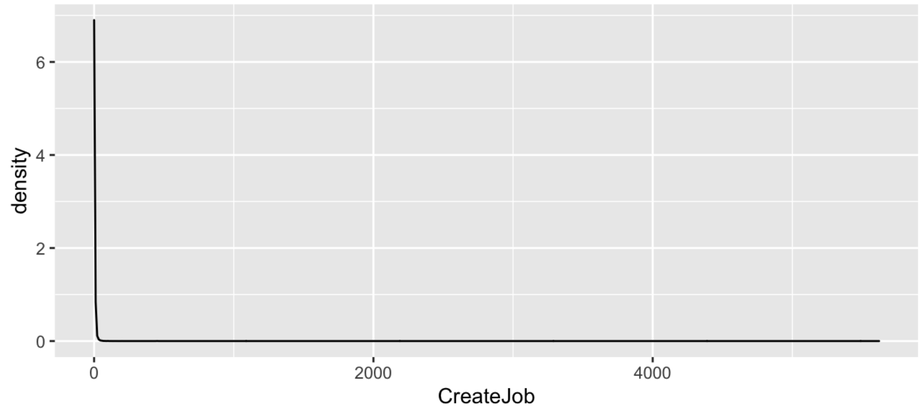
RealEstate

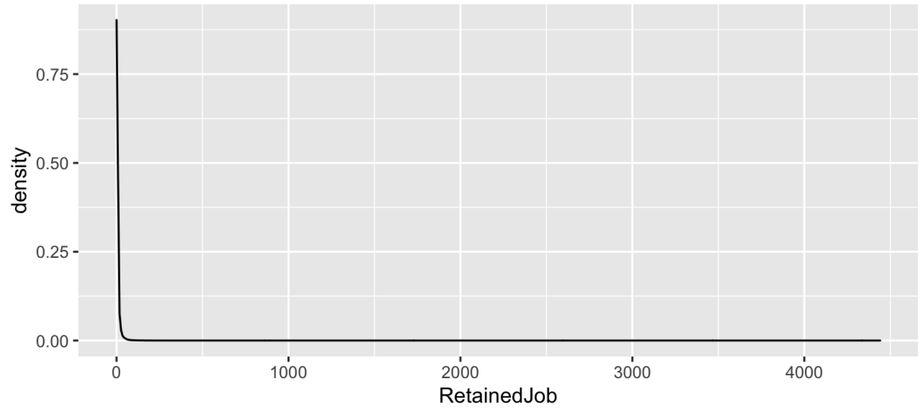
0 =no, 1 = yes

11.06%

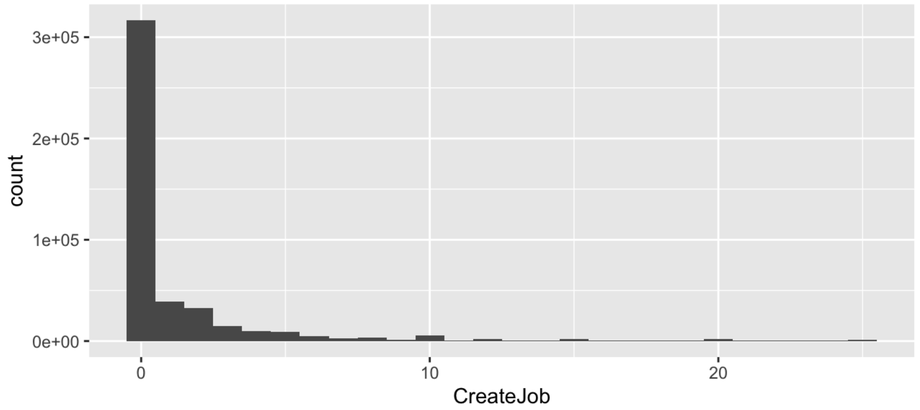


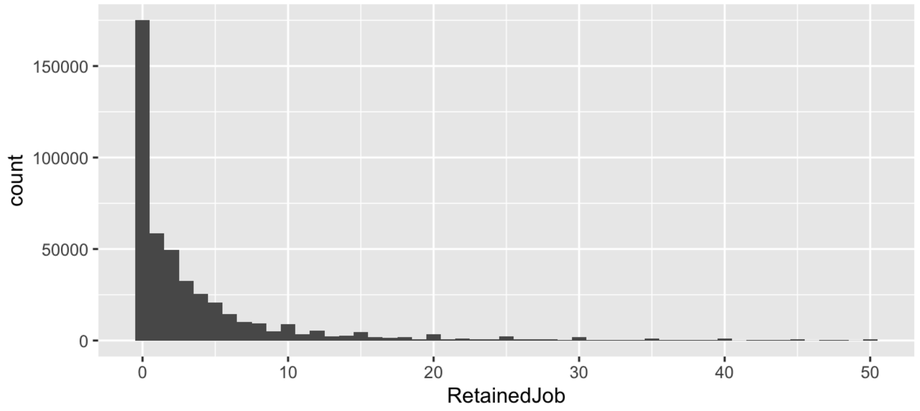
CreateJob	RetainedJob



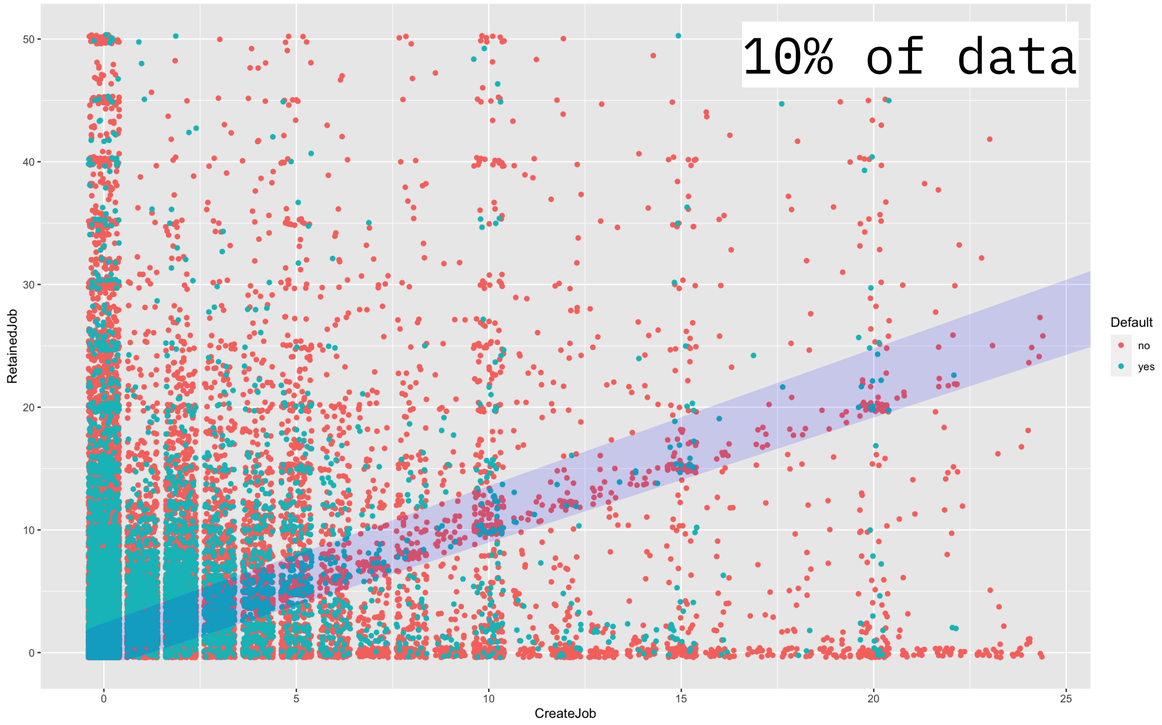


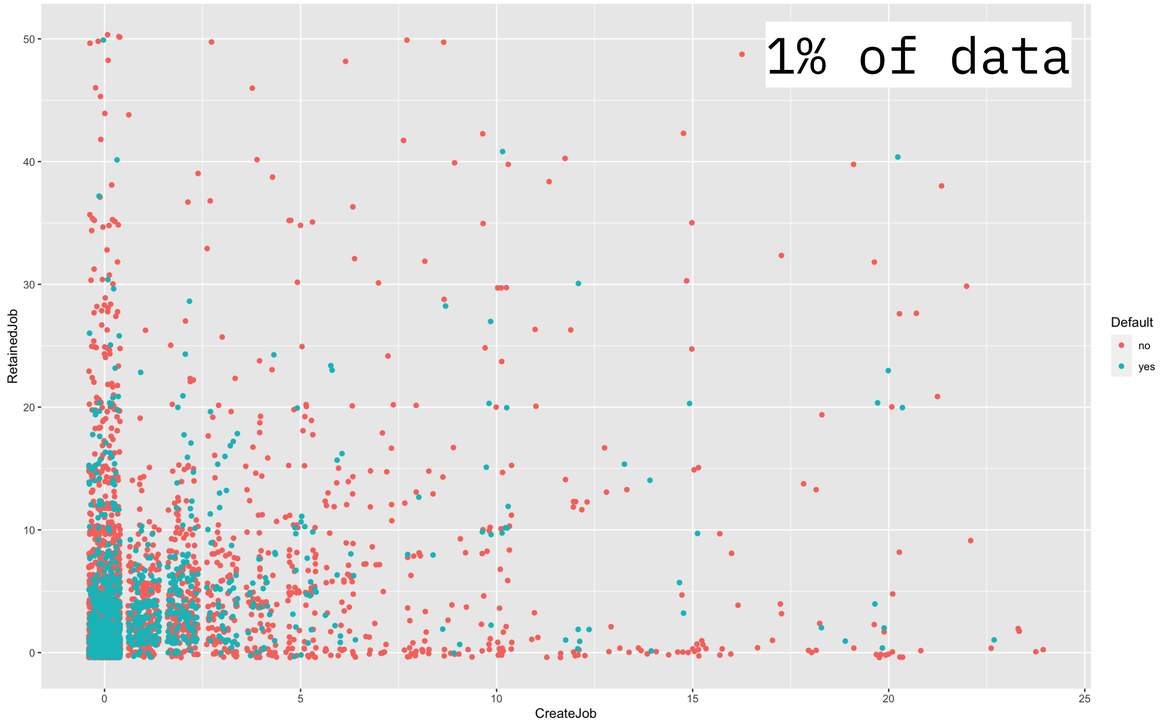
First 99%





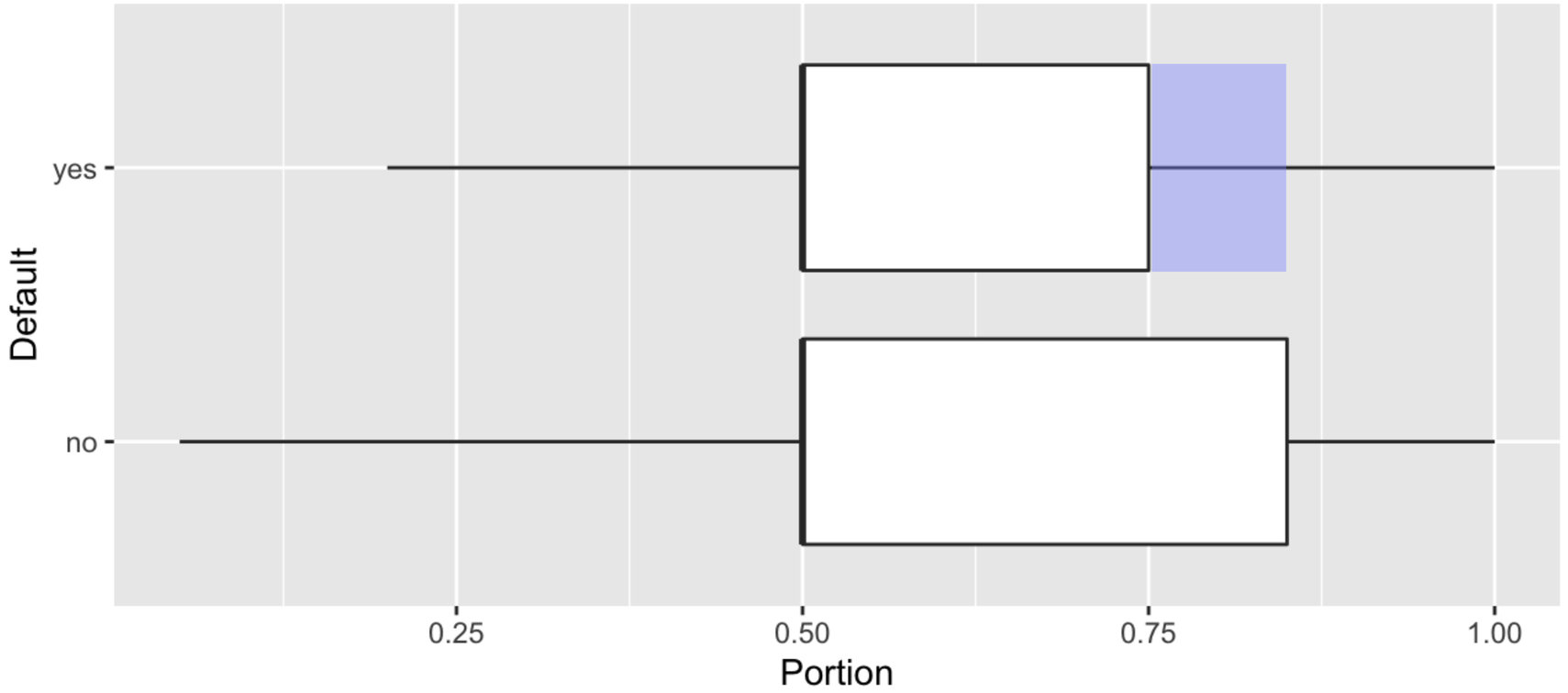
CreateJob	RetainedJob





LOW risk if create = retained **BUT overall is so scattered**

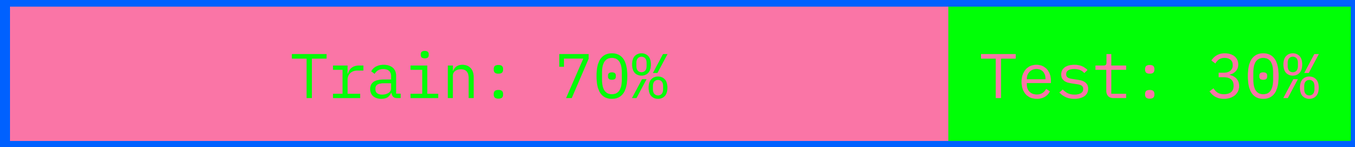
Portion



Model explanation.

Hold out

```
set.seed(007)
```



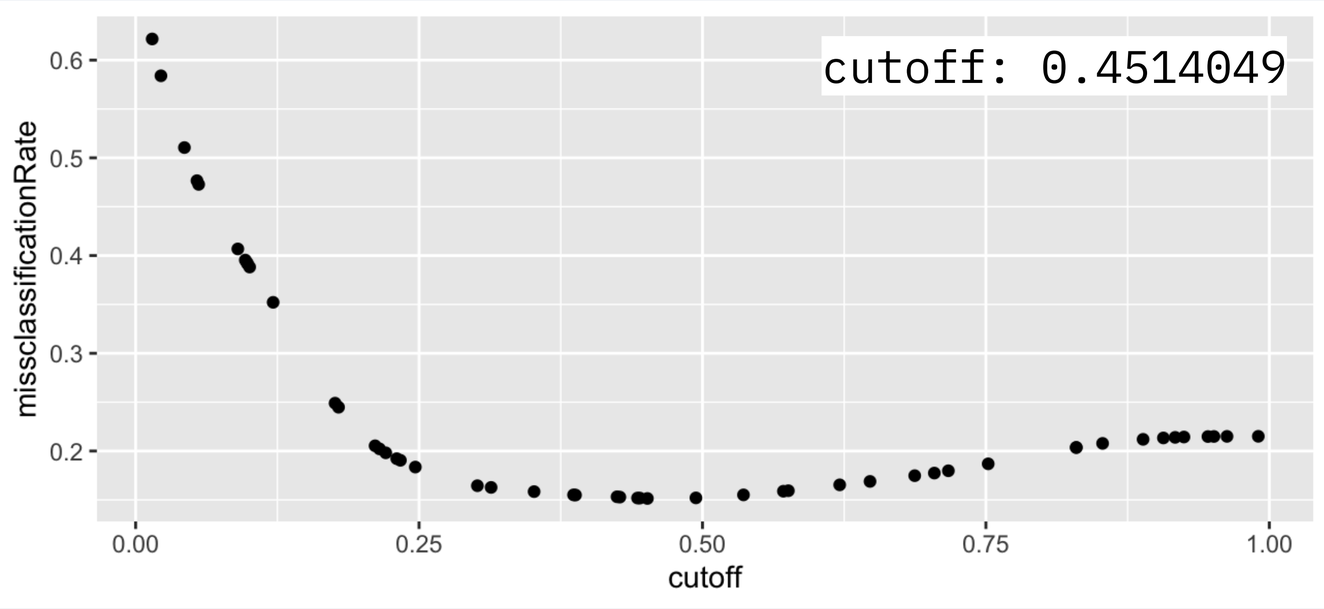
decision_rule()

Function for Logistic regression to return the **best cutoff** value

missclassification: as.character(factor)!=as.character(ref)

```
predict_logistic <- predict(model, test, type = "response")
decision <- decision_rule(test$Default, predict_logistic)
ggplot(decision, aes(cutoff, missclassificationRate)) + geom_point()
cutoff <- decision$cutoff[match(min(decision$missclassificationRate), decision$missclassificationRate)]
```

```
decision_rule <- function(ref, predict_logistic) {

  cutoff <- runif(50)
  missclassificationRate <- c()
  for (i in cutoff) {
    factor <- factor(ifelse(predict_logistic > i, "yes", "no"))
    df <- data.frame(factor = factor, actual = ref)
    n <- nrow(filter(df, as.character(factor)!=as.character(ref)))
    missclassificationRate <- append(missclassificationRate, n/nrow(df))
  }

  return(data.frame(
    cutoff = cutoff,
    missclassificationRate = missclassificationRate
  ))
}
```



Logistic regression

Train

- **Function:** `glm()`
- **Target:** `train$Default`

Predict

- **With:** `test$Default`
- **Decision rule:** `cutoff` from `decision_rule()`
- **Detect to:** `ifelse(predict_logistic > cutoff, "yes", "no")`
- **Positive:** `"no"`

Decision tree

Train

- **Function:** `rpart()`
- **Data:** `train`

Predict

- **With:** `test`
- **Decision rule:** `cutoff from decision_rule()`
- **Positive:** `"no"`

Modeling implementation and evaluation.

Logistic: Predictor ladder

(remove lowest potential predictor one by one until the model outcome does not get better but not a worst one)

Predictor(s)	Precision	Recall
.	0.8792	0.9429
. -State	0.8703	0.9375
. -State -RealEstate	0.8694	0.9397
. -State -RealEstate -DisbursementGross	0.8702	0.9381
. -State -RealEstate -DisbursementGross -LowDoc	0.8689	0.9400
. -State -RealEstate -DisbursementGross -LowDoc -BankInState	0.8635	0.9195
. -State -RealEstate -DisbursementGross -LowDoc -BankInState -NewExist	0.8629	0.9202
. -State -RealEstate -DisbursementGross -LowDoc -BankInState -NewExist -Portion	0.8704	0.9099
. -State -RealEstate -DisbursementGross -LowDoc -BankInState -NewExist -Portion -Recession	0.8806	0.9011

Logistic_I: . -State -RealEstate -DisbursementGross -LowDoc

```
Call:
glm(formula = Default ~ . - State - RealEstate - DisbursementGross -
    LowDoc, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0710  -0.6322  -0.3607  -0.0154   8.4904

Coefficients:
                     Estimate Std. Error  z value Pr(>|z|)
(Intercept)         9.775e-01  6.541e-02   14.945  < 2e-16 ***
NAICS…
Term               -3.454e-02  1.910e-04 -180.912  < 2e-16 ***
NoEmp              -1.723e-02  6.428e-04  -26.802  < 2e-16 ***
NewExist1          -4.666e-02  1.179e-02   -3.958 7.57e-05 ***
CreateJob           3.000e-03  4.005e-04    7.491 6.84e-14 ***
RetainedJob         5.921e-03  7.032e-04    8.419  < 2e-16 ***
RevLineCrY         -4.256e-01  1.261e-02  -33.758  < 2e-16 ***
Franchise1         -8.761e-02  3.257e-02   -2.690 0.007154 **
DaysToDisbursement -2.935e-03  5.758e-05  -50.979  < 2e-16 ***
Recession1          6.265e-01  1.690e-02   37.071  < 2e-16 ***
Portion             4.314e-01  4.228e-02   10.204  < 2e-16 ***
BankInState1       -1.160e+00  1.186e-02  -97.795  < 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: 333115  on 320479  degrees of freedom
Residual deviance: 244813  on 320445  degrees of freedom
AIC: 244883

Number of Fisher Scoring iterations: 7
```

```
Confusion Matrix and Statistics

          Reference
Prediction     no    yes
       no  101786  15844
       yes   5882  13837

               Accuracy : 0.8418
                 95% CI : (0.8399, 0.8437)
    No Information Rate : 0.7839
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4685

 Mcnemar's Test P-Value : < 2.2e-16

              Precision : 0.8653
                 Recall : 0.9454
                     F1 : 0.9036
             Prevalence : 0.7839
         Detection Rate : 0.7411
   Detection Prevalence : 0.8564
      Balanced Accuracy : 0.7058

       'Positive' Class : no
```

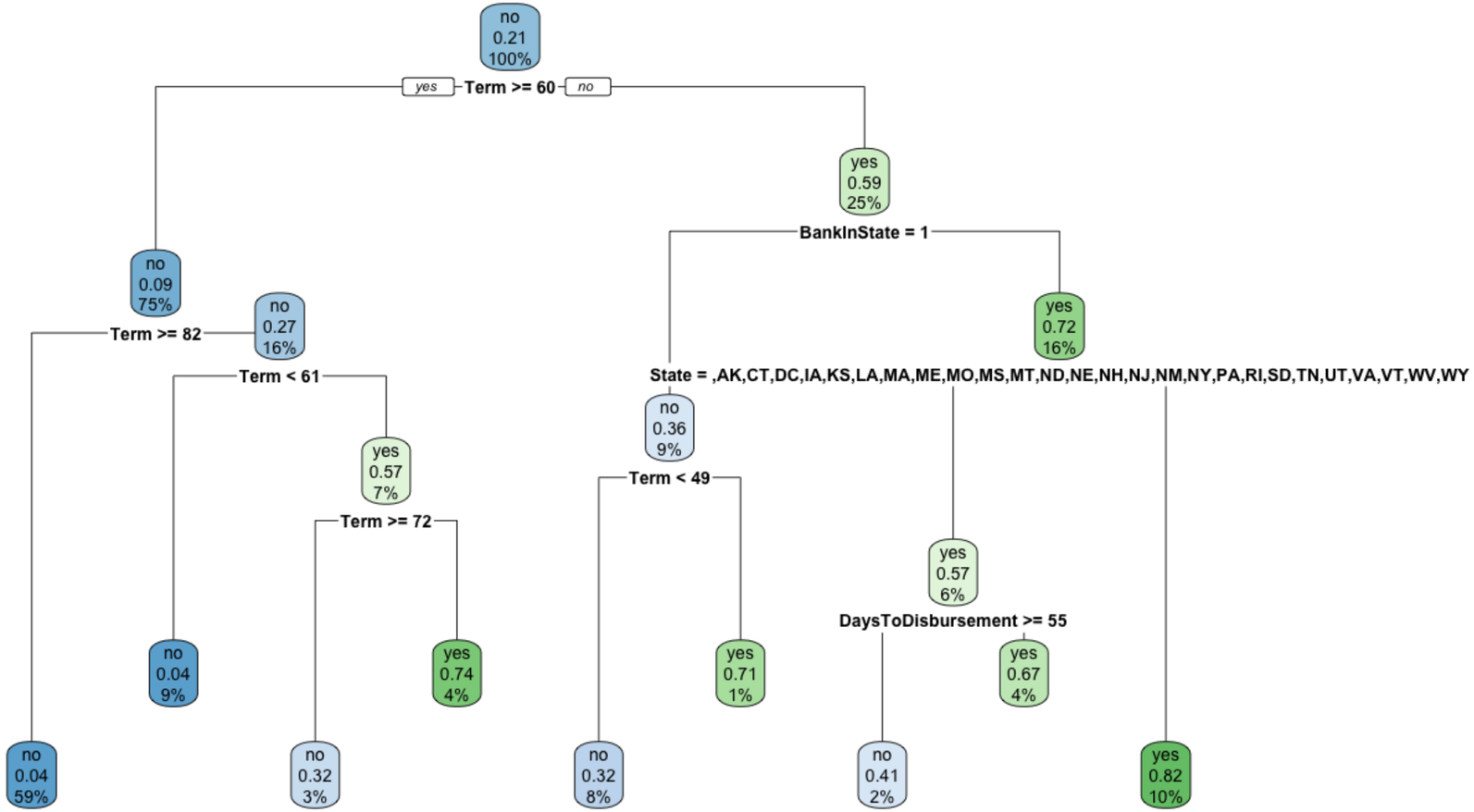
Logistic_II: RealEstate * Portion * Recession

```
Call:
glm(formula = Default ~ Recession * Portion * RealEstate, family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8392  -0.7608  -0.6267  -0.0961   3.2801

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -0.36494    0.01897 -19.239  < 2e-16 ***
Recession1                    3.20652    0.07056  45.446  < 2e-16 ***
Portion                      -1.45361    0.03028 -48.002  < 2e-16 ***
RealEstate1                   1.51182    0.25018   6.043 1.51e-09 ***
Recession1:Portion           -3.96204    0.12459 -31.800  < 2e-16 ***
Recession1:RealEstate1       -3.95879    2.28824  -1.730   0.0836 .
Portion:RealEstate1          -5.06825    0.31952 -15.862  < 2e-16 ***
Recession1:Portion:RealEstate1  4.56192    2.63606   1.731   0.0835 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 333115  on 320479  degrees of freedom
Residual deviance: 308252  on 320472  degrees of freedom
AIC: 308268

Number of Fisher Scoring iterations: 7
```

```
Confusion Matrix and Statistics

          Reference
Prediction     no    yes
       no  104176  25512
       yes   3492   4169

               Accuracy : 0.7888
                 95% CI : (0.7867, 0.791)
    No Information Rate : 0.7839
    P-Value [Acc > NIR] : 4.359e-06

                  Kappa : 0.1477

 Mcnemar's Test P-Value : < 2.2e-16

              Precision : 0.8033
                 Recall : 0.9676
                     F1 : 0.8778
             Prevalence : 0.7839
         Detection Rate : 0.7585
   Detection Prevalence : 0.9442
      Balanced Accuracy : 0.5540

       'Positive' Class : no
```

Rejected

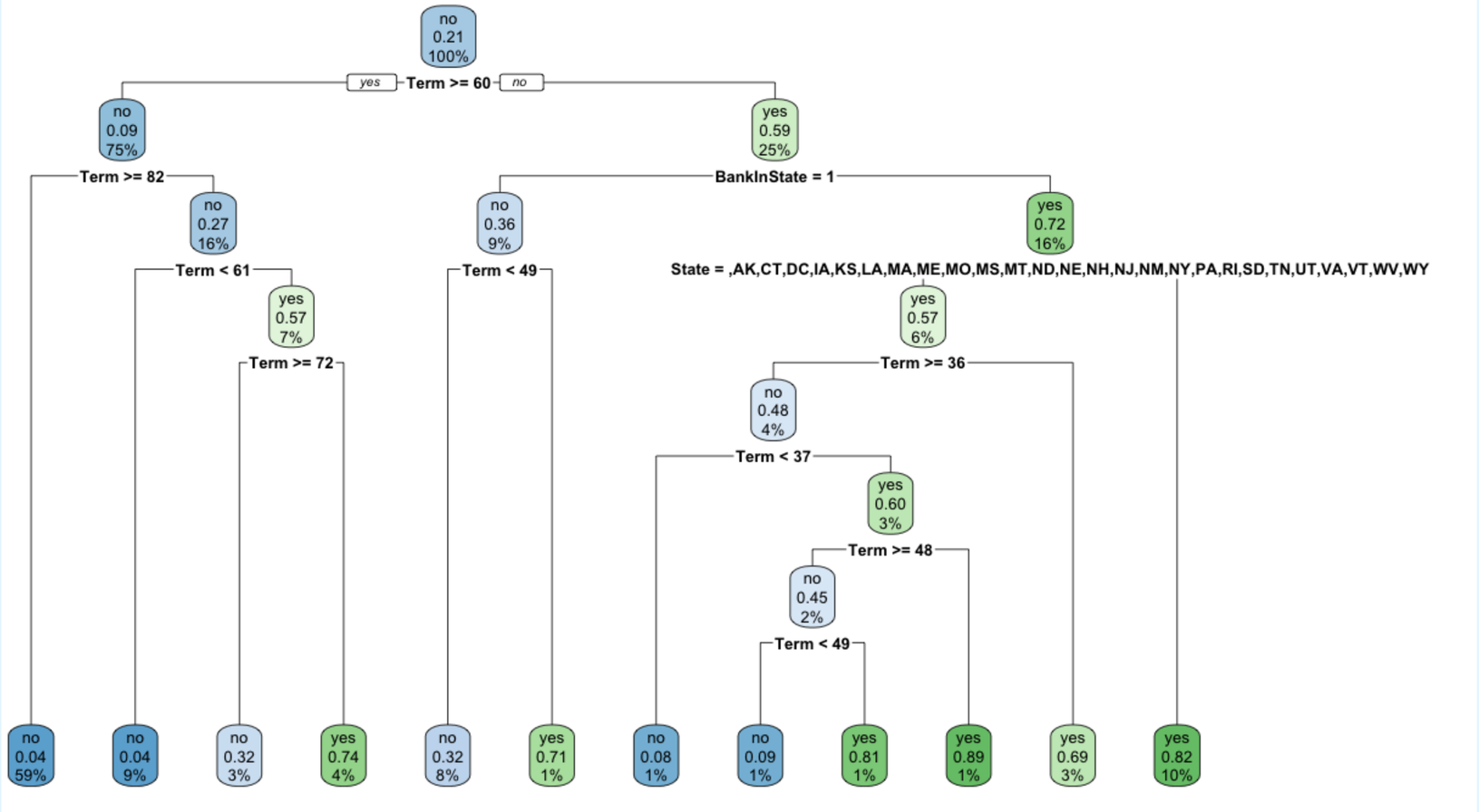
Decision tree_I: .




```
> tree$variable.importance
            Term        BankInState              State            Portion
     45174.497187        5515.990840        3718.069219        1393.091612
      RetainedJob          Recession DaysToDisbursement   DisbursementGross
      1249.610085        1141.580905         787.807091         387.743257
       CreateJob             LowDoc              NAICS              NoEmp
       364.613163           7.353862           4.377319           3.199073
```

```
Confusion Matrix and Statistics

          Reference
Prediction     no    yes
       no  101524   9672
       yes   6144  20009

               Accuracy : 0.8848
                 95% CI : (0.8831, 0.8865)
    No Information Rate : 0.7839
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6448

 Mcnemar's Test P-Value : < 2.2e-16

              Precision : 0.9130
                 Recall : 0.9429
                     F1 : 0.9277
             Prevalence : 0.7839
         Detection Rate : 0.7392
   Detection Prevalence : 0.8096
      Balanced Accuracy : 0.8085

       'Positive' Class : no
```

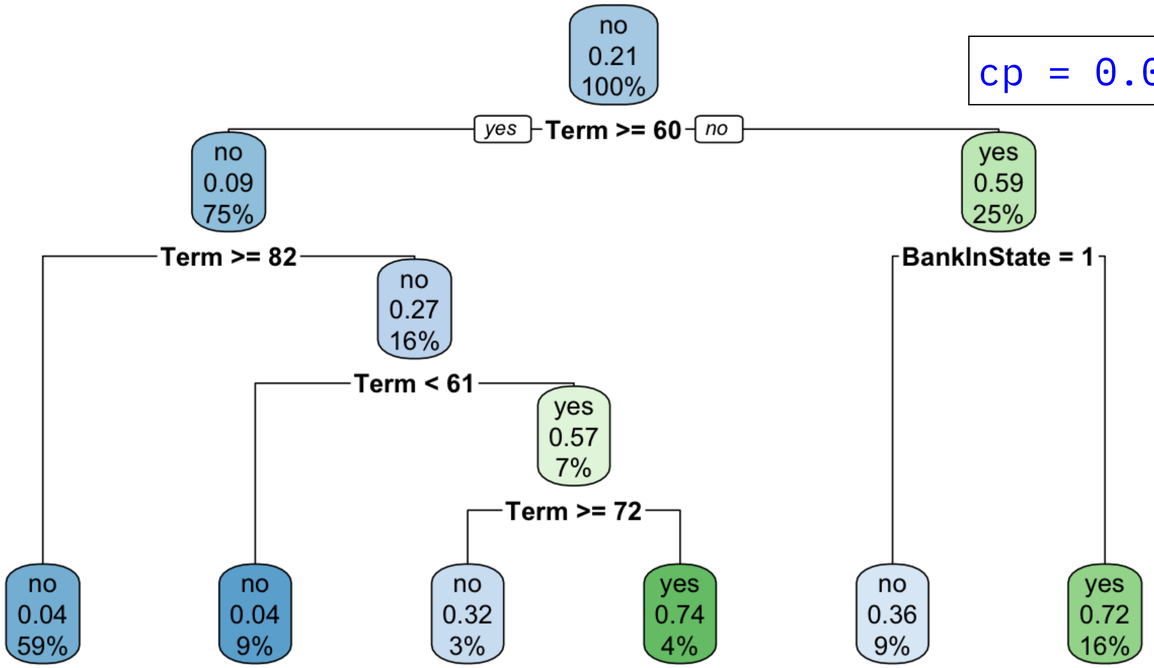
Decision tree_II: Top 4 **tree$variable.importance** of .



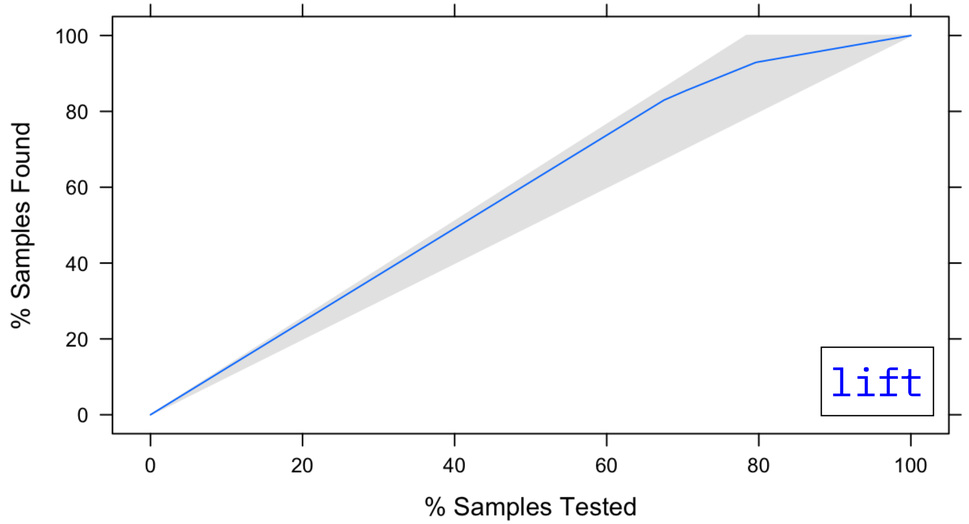

```
> tree$variable.importance
      Term BankInState        State      Portion
 49014.558    5515.991    4087.394    1484.059
```

```
Confusion Matrix and Statistics

          Reference
Prediction     no    yes
       no  101824   8513
       yes   5844  21168

               Accuracy : 0.8955
                 95% CI : (0.8938, 0.8971)
    No Information Rate : 0.7839
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6811

 Mcnemar's Test P-Value : < 2.2e-16

              Precision : 0.9228
                 Recall : 0.9457
                     F1 : 0.9341
             Prevalence : 0.7839
         Detection Rate : 0.7414
   Detection Prevalence : 0.8033
      Balanced Accuracy : 0.8295

       'Positive' Class : no
```

Cross validation (n = 10): Decision tree _I and _II

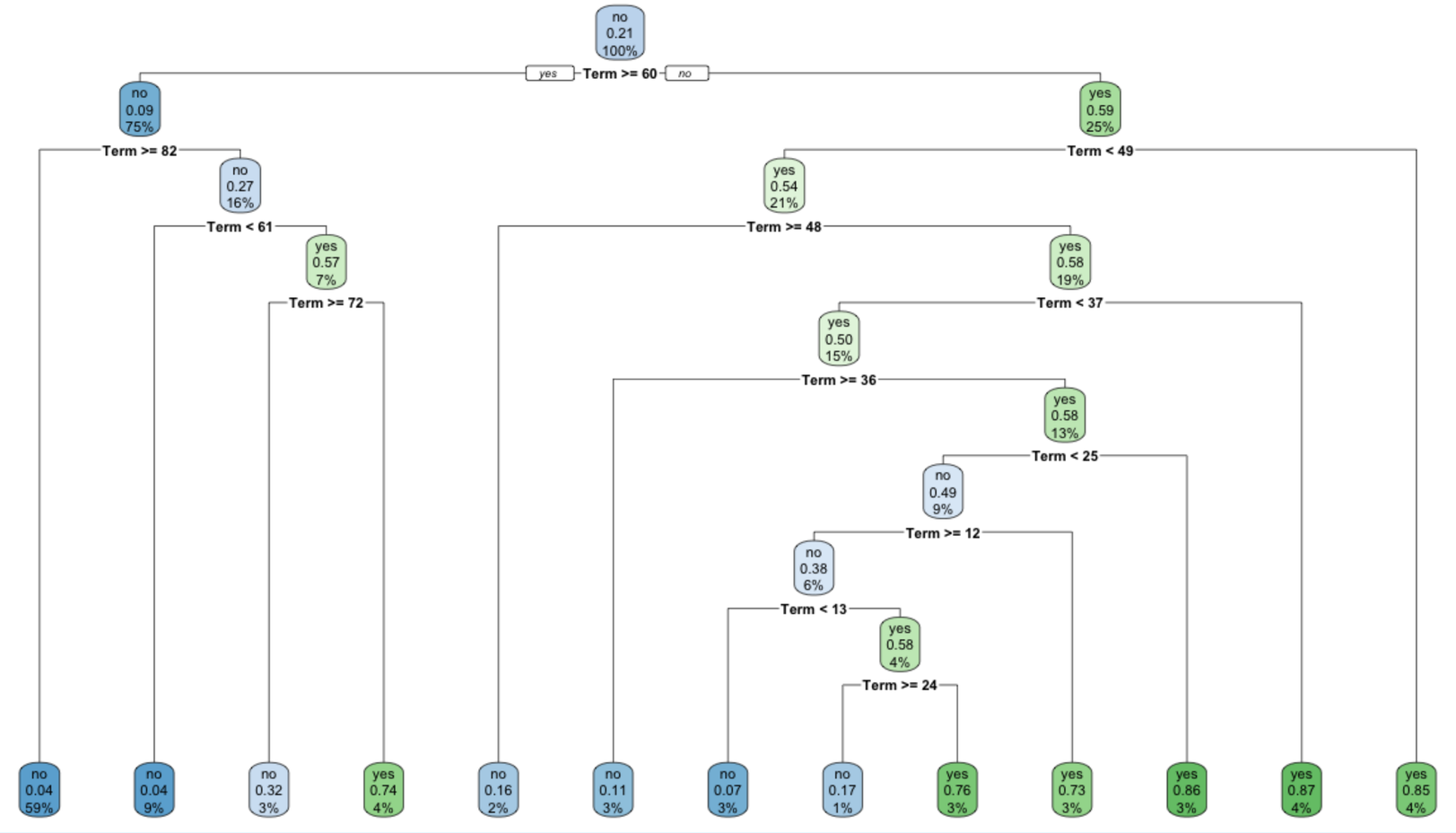


cp = 0.03063734



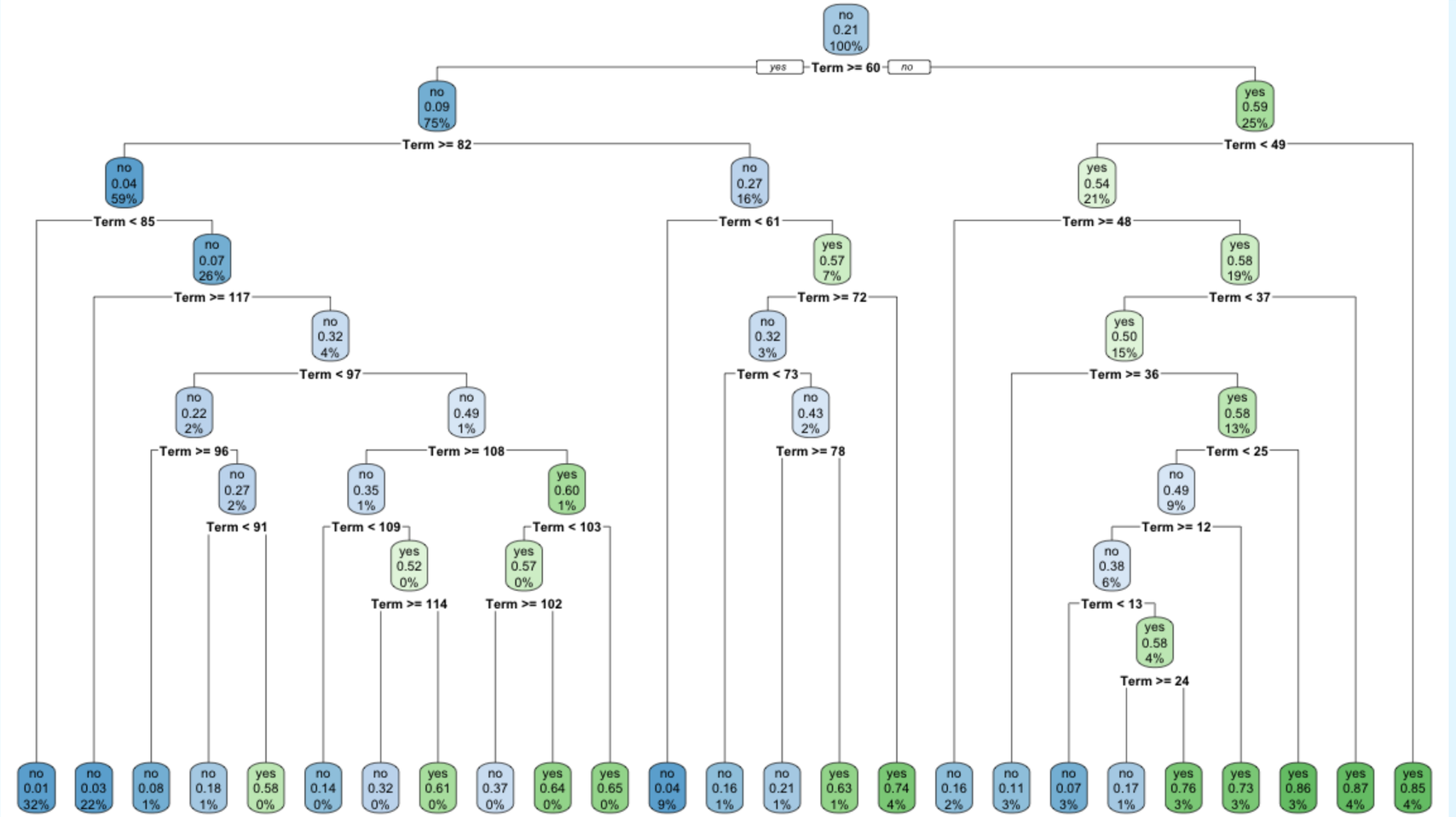

```
Confusion Matrix and Statistics

          Reference
Prediction     no    yes
       no  100049   9313
       yes   7619  20368

               Accuracy : 0.8767
                 95% CI : (0.875, 0.8785)
    No Information Rate : 0.7839
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6285

 Mcnemar's Test P-Value : < 2.2e-16

              Precision : 0.9148
                 Recall : 0.9292
                     F1 : 0.9220
             Prevalence : 0.7839
         Detection Rate : 0.7284
   Detection Prevalence : 0.7962
      Balanced Accuracy : 0.8077

       'Positive' Class : no
```

Decision tree_III: Top 1 **tree$variable.importance** of all time




```
Confusion Matrix and Statistics

          Reference
Prediction     no    yes
       no  102037   6071
       yes   5631  23610

               Accuracy : 0.9148
                 95% CI : (0.9133, 0.9163)
    No Information Rate : 0.7839
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7472

 Mcnemar's Test P-Value : 4.945e-05

              Precision : 0.9438
                 Recall : 0.9477
                     F1 : 0.9458
             Prevalence : 0.7839
         Detection Rate : 0.7429
   Detection Prevalence : 0.7871
      Balanced Accuracy : 0.8716

       'Positive' Class : no
```

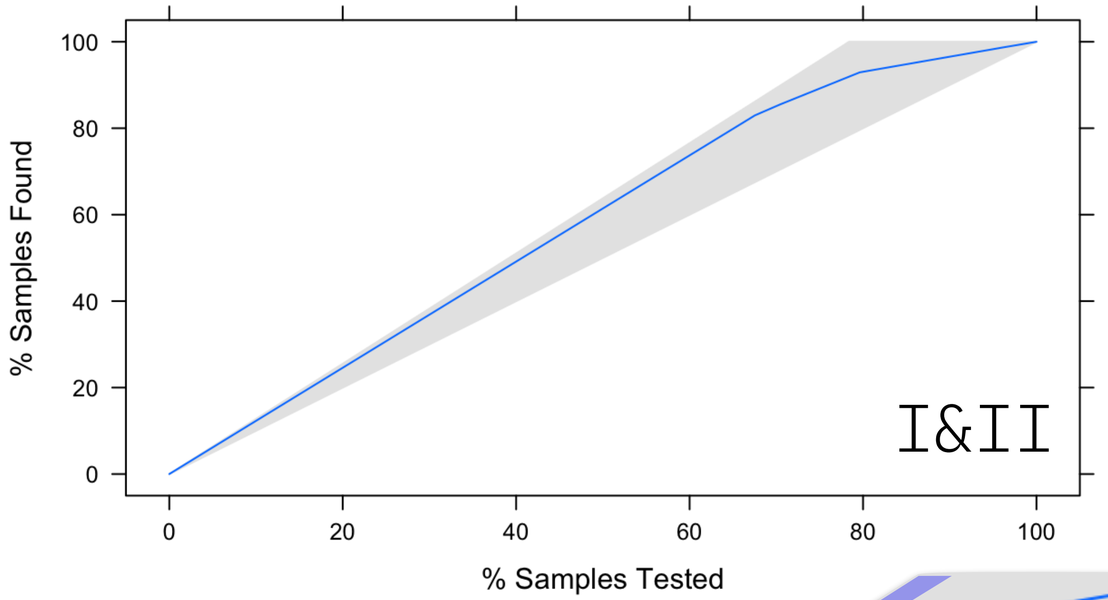
Cross validation Decision tree_III

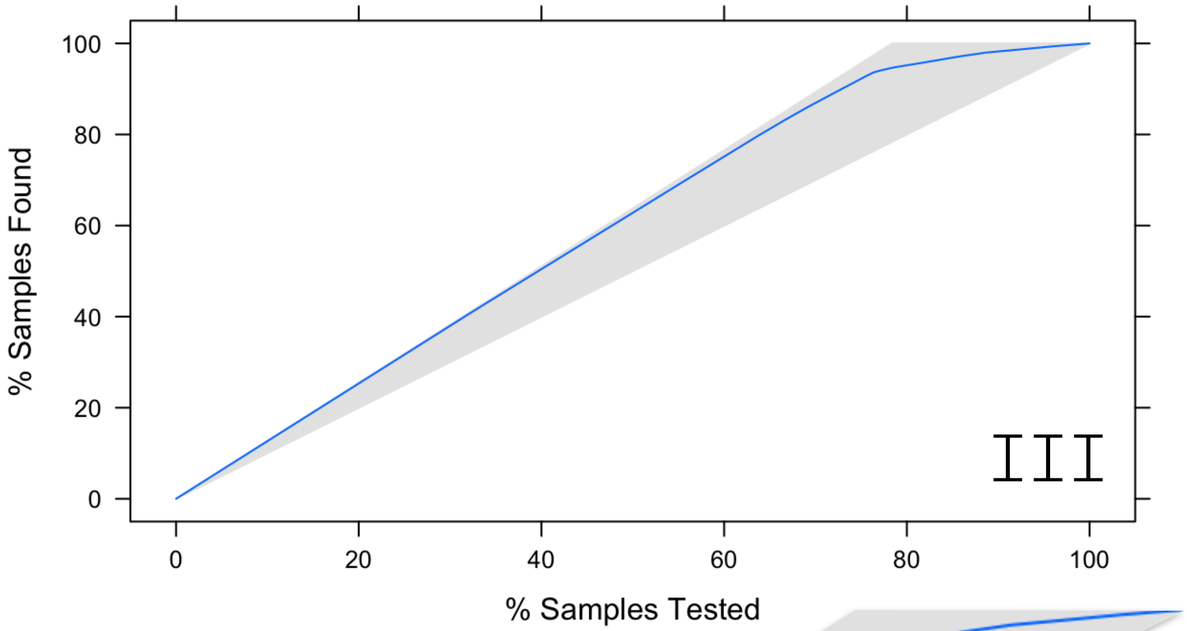


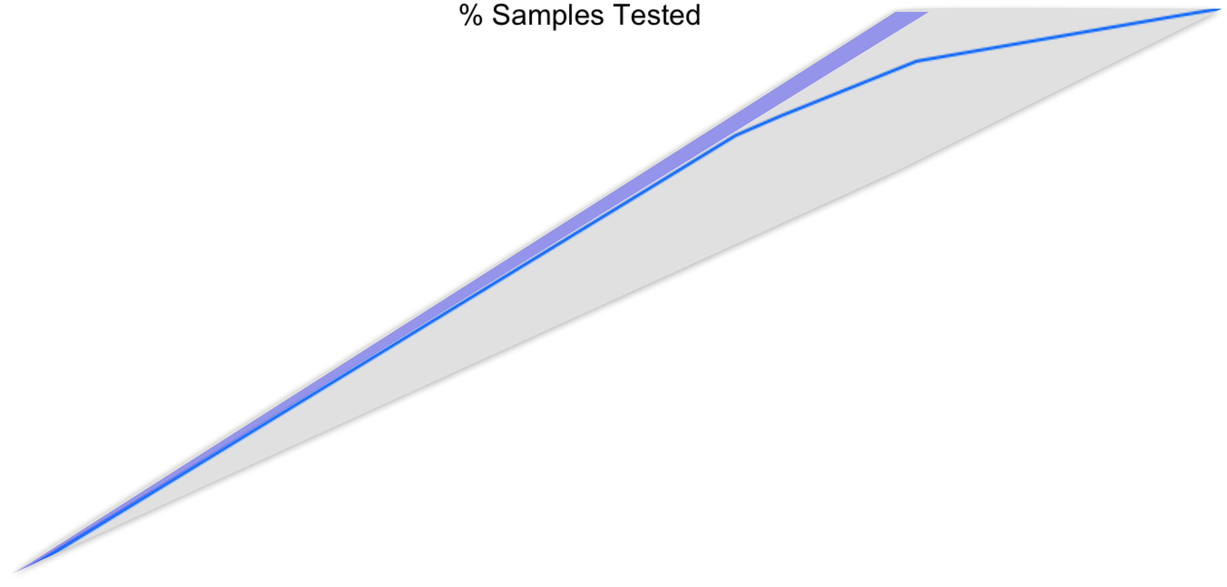
cp = 0.0006030487

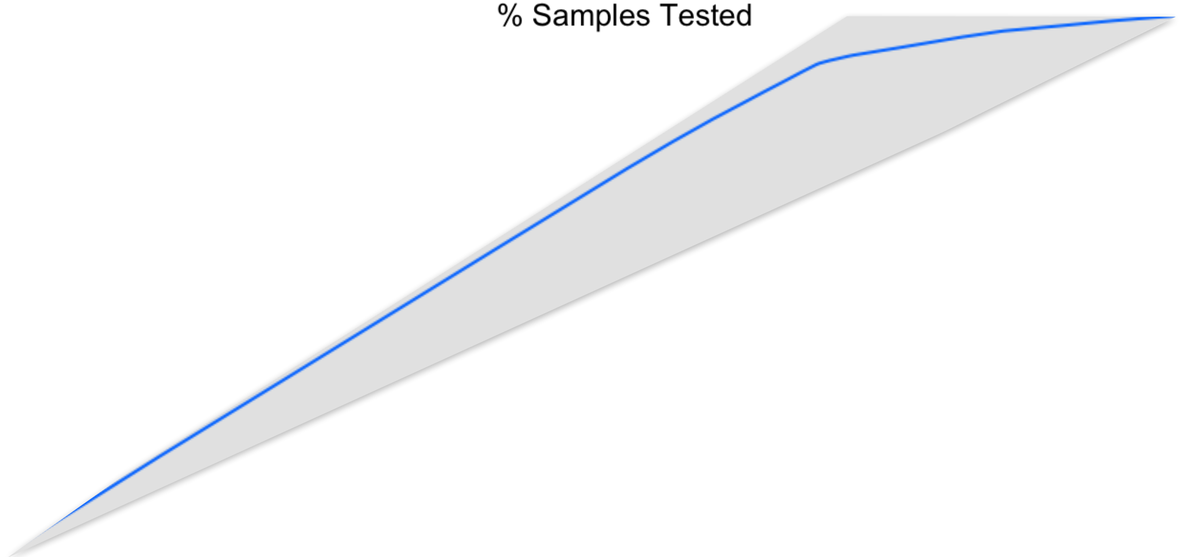
```
Confusion Matrix and Statistics

          Reference
Prediction     no    yes
       no  100869   4127
       yes   6799  25554

               Accuracy : 0.9205
                 95% CI : (0.919, 0.9219)
    No Information Rate : 0.7839
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7726

 Mcnemar's Test P-Value : < 2.2e-16

            Precision : 0.9607
               Recall : 0.9369
                   F1 : 0.9486
           Prevalence : 0.7839
       Detection Rate : 0.7344
 Detection Prevalence : 0.7644
    Balanced Accuracy : 0.8989

       'Positive' Class : no
```

Lift: Decision tree_(I&II) and _III



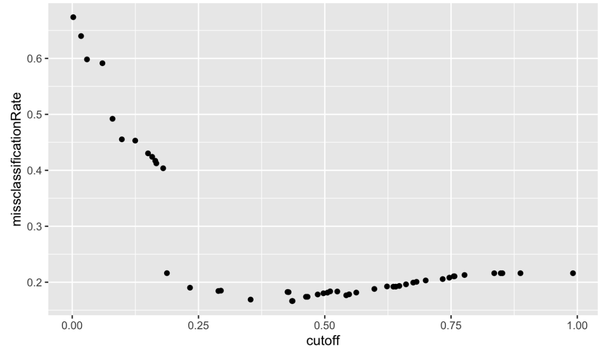






Logistic_III: Top 4 **tree$variable.importance** of . Except `State`

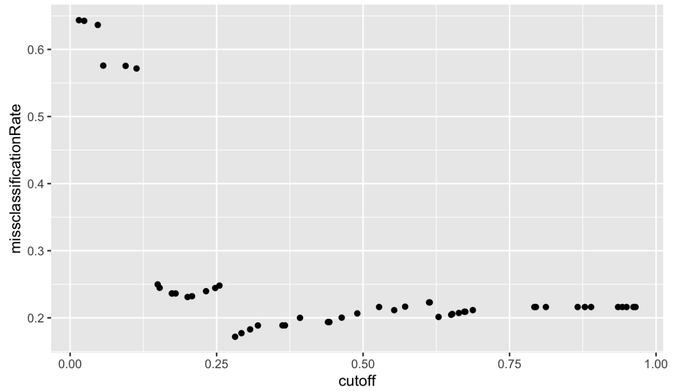
```
Call:
glm(formula = Default ~ Term + BankInState + Portion, family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9216  -0.6375  -0.3953  -0.0286   4.7255

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.9683593  0.0210439   46.02   <2e-16 ***
Term        -0.0339405  0.0001795 -189.06   <2e-16 ***
BankInState1 -1.2160535  0.0116539 -104.35   <2e-16 ***
Portion      0.7845532  0.0339028   23.14   <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: 333115  on 320479  degrees of freedom
Residual deviance: 257593  on 320476  degrees of freedom
AIC: 257601

Number of Fisher Scoring iterations: 6
```




```
Confusion Matrix and Statistics

          Reference
Prediction     no    yes
       no  101117  16302
       yes   6551  13379

               Accuracy : 0.8336
                 95% CI : (0.8316, 0.8356)
    No Information Rate : 0.7839
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4426

 Mcnemar's Test P-Value : < 2.2e-16

              Precision : 0.8612
                 Recall : 0.9392
                     F1 : 0.8985
             Prevalence : 0.7839
         Detection Rate : 0.7362
   Detection Prevalence : 0.8549
      Balanced Accuracy : 0.6950

       'Positive' Class : no
```

Logistic_IV: Default ~ Term

```
Call:
glm(formula = Default ~ Term, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5637  -0.5615  -0.5615  -0.0458   4.5846

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.8737243  0.0107655   81.16   <2e-16 ***
Term        -0.0314443  0.0001656 -189.90   <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: 333115  on 320479  degrees of freedom
Residual deviance: 269871  on 320478  degrees of freedom
AIC: 269875

Number of Fisher Scoring iterations: 6
```




```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  93837  9763
       yes 13831 19918

               Accuracy : 0.8282
                 95% CI : (0.8262, 0.8302)
    No Information Rate : 0.7839
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5169

 Mcnemar's Test P-Value : < 2.2e-16

              Precision : 0.9058
                 Recall : 0.8715
                     F1 : 0.8883
             Prevalence : 0.7839
         Detection Rate : 0.6832
   Detection Prevalence : 0.7543
      Balanced Accuracy : 0.7713

       'Positive' Class : no
```

Discussion and conclusion.

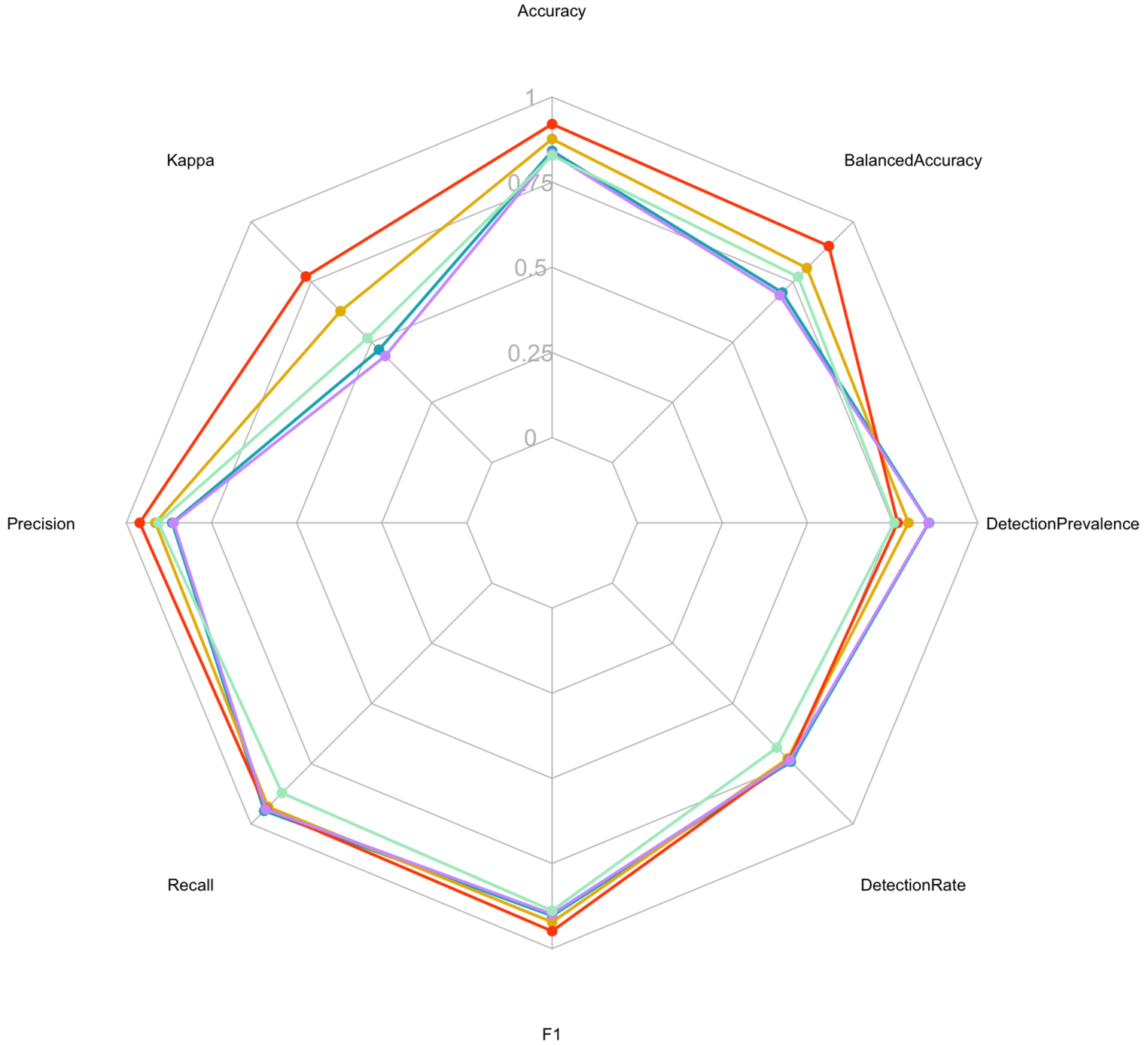
Summarizing all model

Model	Accuracy	Kappa	Precision	Recall	F1	Detection Rate	Detection Prevalence	Balanced Accuracy
Ladder	0.8418	0.4685	0.8653	0.9454	0.9036	0.7411	0.8564	0.7058
Cross I&II	0.8767	0.6285	0.9148	0.9292	0.9220	0.7284	0.7962	0.8077
Cross III	0.9205	0.7726	0.9607	0.9369	0.9486	0.7344	0.7644	0.8989
glm ~top4	0.8336	0.4426	0.8612	0.9392	0.8985	0.7362	0.8549	0.6950
glm ~Term	0.8282	0.5169	0.9058	0.8715	0.8883	0.6832	0.7543	0.7713

Prevalence = No Information Rate = **0.7839**

Summarizing all model

Ladder
Cross I&II
Cross III
glm ~top4
glm ~Term



Selected model: Cross III

rpart(Default ~ **Term**, train, cp = 0.0006030487)

```
Confusion Matrix and Statistics

          Reference
Prediction     no    yes
       no  100869   4127
       yes   6799  25554

               Accuracy : 0.9205
                 95% CI : (0.919, 0.9219)
    No Information Rate : 0.7839
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7726

 Mcnemar's Test P-Value : < 2.2e-16

              Precision : 0.9607
                 Recall : 0.9369
                     F1 : 0.9486
             Prevalence : 0.7839
         Detection Rate : 0.7344
   Detection Prevalence : 0.7644
      Balanced Accuracy : 0.8989

       'Positive' Class : no
```

Prediction	Referenced no	Referenced yes
no	100869	4127
yes	6799	25554

0.9607	0.9369	0.9486
Precision	**Recall**	**F1**

Thank You.