Central University of Finance and Economics

中央财经大学

经济统计学课程

代码文档

吴宇翀

2017310836

WUYUCHONG.COM

指导老师：王文静

2020 年 6 月 21 日

目录

1 环境配置与程序包

```
knitr::opts_chunk$set(fig.pos = 'H', warning = FALSE, message = FALSE)
library(knitr)
library(tidyverse)
library(caret)
library(kernlab)
library(pROC)
library(kableExtra)
# base_family = 'STXihei'
```

2 数据集与描述分析

2.1 数据集

```
dat = read.csv("data.csv", header = TRUE)
dat = dat[2:ncol(dat)]

dat_complete = dat %>%
  mutate(Terminate = 0)
dat_complete$Terminate[which(dat_complete$EmploymentStatus %in% c("Voluntarily Terminated", "Termi
dat_complete$Terminate = as.factor(dat_complete$Terminate)

training = dat_complete
dat_complete = dat_complete %>%
  filter(Department != "Executive Office")

table = read.csv("dictionary.csv", header = TRUE)
names(table) = c(" 变量名", " 变量描述", " 数据格式")
kable(table, booktabs=TRUE, format="latex", caption = " 变量解释和类型") %>%
  kable_styling(latex_options=c("scale_down", "HOLD_position"))
```

表 1: 变量解释和类型

变量名	变量描述	数据格式
Employee Name	Employee’ s full name	Text
EmpID	Employee ID is unique to each employee	Text
MarriedID	Is the person married (1 or 0 for yes or no)	Binary
MaritalStatusID	Marital status code that matches the text field MaritalDesc	Integer
EmpStatusID	Employment status code that matches text field EmploymentStatus	Integer
DeptID	Department ID code that matches the department the employee works in	Integer
PerfScoreID	Performance Score code that matches the employee’ s most recent performance score	Integer
FromDiversityJobFairID	Was the employee sourced from the Diversity job fair? 1 or 0 for yes or no	Binary
PayRate	The person’ s hourly pay rate. All salaries are converted to hourly pay rate	Float
Termd	Has this employee been terminated - 1 or 0	Binary
PositionID	An integer indicating the person’ s position	Integer
Position	The text name/title of the position the person has	Text
State	The state that the person lives in	Text
Zip	The zip code for the employee	Text
DOB	Date of Birth for the employee	Date
Sex	Sex - M or F	Text
MaritalDesc	The marital status of the person (divorced, single, widowed, separated, etc)	Text
CitizenDesc	Label for whether the person is a Citizen or Eligible NonCitizen	Text
HispanicLatino	Yes or No field for whether the employee is Hispanic/Latino	Text
RaceDesc	Description/text of the race the person identifies with	Text
DateofHire	Date the person was hired	Date
DateofTermination	Date the person was terminated, only populated if, in fact, Termd = 1	Date
TermReason	A text reason / description for why the person was terminated	Text
EmploymentStatus	A description/category of the person’ s employment status. Anyone currently working full time = Active	Text
Department	Name of the department that the person works in	Text
ManagerName	The name of the person’ s immediate manager	Text
ManagerID	A unique identifier for each manager.	Integer
RecruitmentSource	The name of the recruitment source where the employee was recruited from	Text
PerformanceScore	Performance Score text/category (Fully Meets, Partially Meets, PIP, Exceeds)	Text
EngagementSurvey	Results from the last engagement survey, managed by our external partner	Float
EmpSatisfaction	A basic satisfaction score between 1 and 5, as reported on a recent employee satisfaction survey	Integer
SpecialProjectsCount	The number of special projects that the employee worked on during the last 6 months	Integer
LastPerformanceReviewDate	The most recent date of the person’ s last performance review.	Date
DaysLateLast30	The number of times that the employee was late to work during the last 30 days	Integer

2.2 描述分析

2.2.1 时薪（美元）

```
ggplot(dat_complete, aes(x = PayRate, fill = Terminate)) +
  geom_density(alpha = 0.3) +
  theme_minimal() +
  scale_fill_manual(values = c("#037418", "darkred"))
```

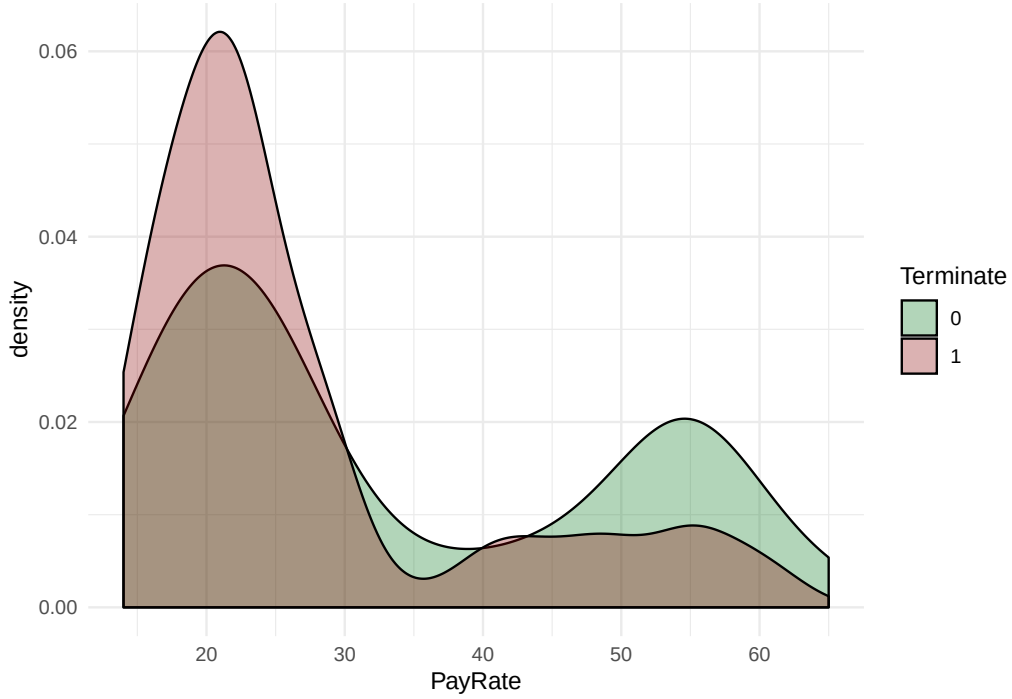

图 1: 离职与在职两类员工日薪分布密度图（红色代表离职）

```
ggplot(dat_complete, aes(x = Sex, y = PayRate, fill = as.factor(MarriedID))) +
  geom_violin(alpha = 0.3) +
  theme_minimal() +
  labs(fill = "Married")
```

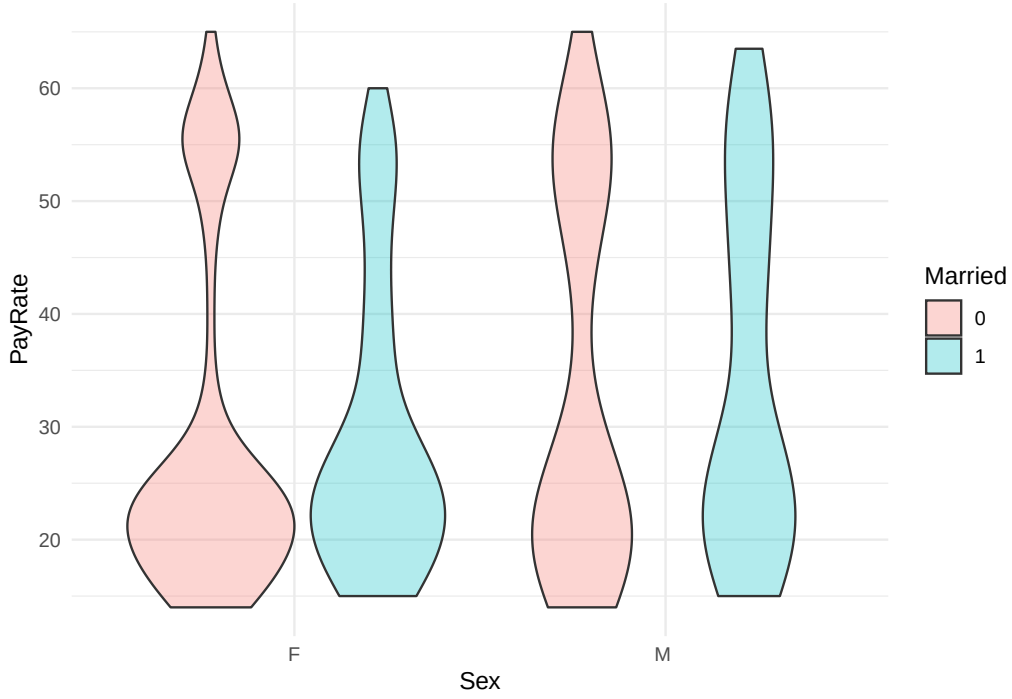

图 2: 不同性别员工日薪分布密度图（蓝色代表已婚）

2.2.2 参与感

```
dat_complete %>%
  ggplot(mapping = aes(x = reorder(Department, EngagementSurvey), y = EngagementSurvey, fill = Ter
    geom_boxplot(alpha = 0.5) +
    labs(x = "Department", y = "Rate for Engagement", fill = "Terminate") +
    theme_minimal() +
    scale_fill_manual(values = c("#037418", "darkred"))
```

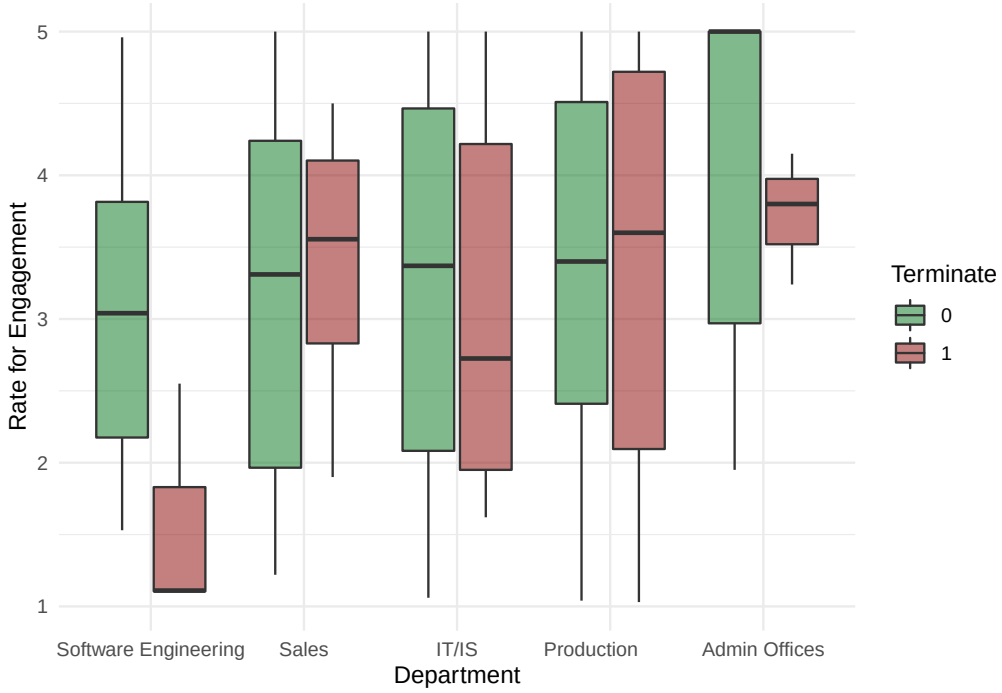

图 3: 离职与在职两类不同部门员工参与感箱线图（红色代表离职）

2.2.3 绩效

```
ggplot(dat_complete, aes(x = EmploymentStatus, fill = PerformanceScore)) +
  geom_bar(stat = "count", position = "fill") +
  theme_minimal() +
  labs(y = "count") +
  coord_flip()
```

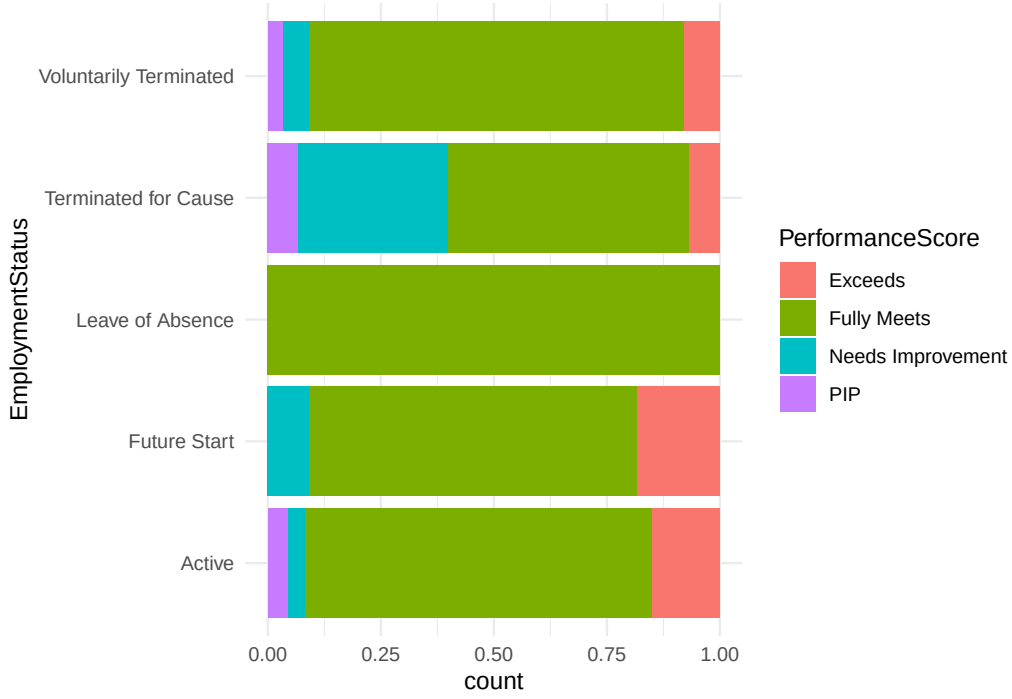

图 4: 不同任职状况的员工绩效（红色代表离职）

3 建立解释模型

3.1 拟合

```
logit2 = glm(Terminate ~ Sex + MaritalDesc + Department + PerformanceScore + EngagementSurvey + Em
logit2_sum = summary(logit2)
kable(logit2_sum$coefficients, format="latex", booktabs=TRUE, caption = "Logit 回归系数表", digit =
  kable_styling(latex_options=c("HOLD_position"))
```

表 2: Logit 回归系数表

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.70	1.52	0.46	0.65
SexM	-0.04	0.27	-0.16	0.88
MaritalDescMarried	-0.57	0.43	-1.33	0.18
MaritalDescSeparated	-2.60	1.13	-2.30	0.02
MaritalDescSingle	-1.15	0.43	-2.67	0.01
MaritalDescWidowed	0.08	0.85	0.09	0.93
DepartmentIT/IS	1.32	1.21	1.09	0.28
DepartmentProduction	-1.05	1.10	-0.96	0.34
DepartmentSales	-2.10	1.22	-1.72	0.09
DepartmentSoftware Engineering	1.51	1.22	1.24	0.22
PerformanceScoreFully Meets	0.77	0.45	1.73	0.08
PerformanceScoreNeeds Improvement	1.61	0.65	2.47	0.01
PerformanceScorePIP	1.14	0.84	1.36	0.17
EngagementSurvey	0.02	0.10	0.21	0.83
EmpSatisfaction	0.06	0.16	0.38	0.71
SpecialProjectsCount	-0.52	0.28	-1.91	0.06
PayRate	-0.01	0.02	-0.90	0.37

3.2 预测

```
set.seed(1)
inTraining <- createDataPartition(dat_complete$Terminate, p = .75, list = FALSE)
train <- dat_complete[inTraining,]
test <- dat_complete[-inTraining,]
```

```
logit3 = glm(Terminate ~ Sex + MaritalDesc + Department + PerformanceScore + EngagementSurvey + Em
probability = predict(logit3, test, type = "response")
distribution = as.data.frame(probability)
distribution = cbind(distribution, group = test$Terminate)
ggplot(distribution, aes(x = probability, fill = group)) +
  geom_density(alpha = 0.3) +
  theme_minimal() +
  scale_fill_manual(values = c("#037418", "darkred"))
```

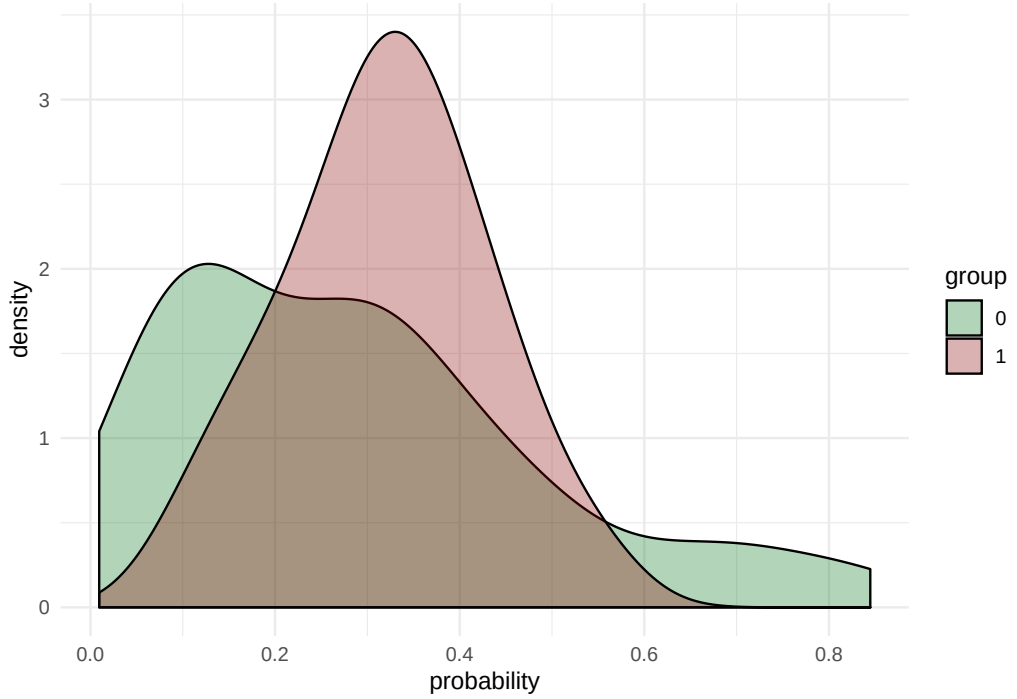

图 5: 预测的离职概率值（红色代表已知为离职）

```
testPred = probability
testPred[testPred > 0.5] = 1
testPred[testPred <= 0.5] = 0
testPred = as.factor(testPred)
```

3.3 混淆矩阵与验证结果

```
confusion = confusionMatrix(data = testPred,
                reference = test$Terminate,
                positive = "1")
kable(as.data.frame(confusion$table), format="latex", booktabs=TRUE, caption = " 混淆矩阵表") %>%
  kable_styling(latex_options=c("HOLD_position"))
```

表 3: 混淆矩阵表

Prediction	Reference	Freq
0	0	44
1	0	7
0	1	24
1	1	1

```
table = as.data.frame(confusion$overall)
names(table) = c(" 指标值")
table = t(table)
rownames(table) = NULL
kable(table, booktabs=TRUE, format="latex", caption = " 验证结果表", digit = 3) %>%
  kable_styling(latex_options=c("HOLD_position"))
```

表 4: 验证结果表

Accuracy	Kappa	AccuracyLower	AccuracyUpper	AccuracyNull	AccuracyPValue	McnemarPValue
0.592	-0.118	0.473	0.704	0.671	0.942	0.004

```
table = as.data.frame(confusion$byClass[1:5])
names(table) = c(" 指标值")
table = t(table)
kable(table, format="latex", booktabs=TRUE, caption = " 灵敏度和特异度等指标表", digit = 3) %>%
  kable_styling(latex_options=c("HOLD_position"))
```

表 5: 灵敏度和特异度等指标表

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision
指标值	0.04	0.863	0.125	0.647	0.125

3.4 接受者操作特征（ROC）曲线

```
rocCurve = roc(response = test$Terminate,
               predictor = probability,
               levels = rev(levels(test$Terminate)),
```

```
            plot = TRUE,
            print.thres=TRUE, print.auc=TRUE)
```

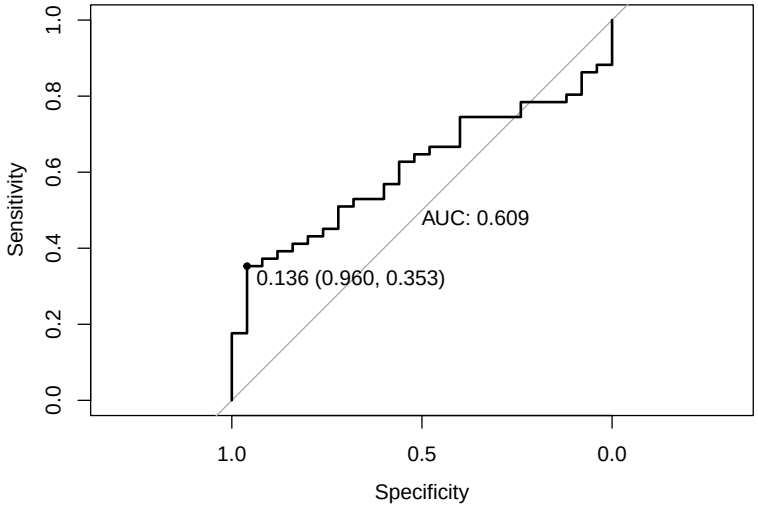

图 6: Logit 模型的 ROC 曲线

4 预测模型的选择

4.1 抽样、训练与评价指标

```
set.seed(1)
fitControl <- trainControl(## 10-fold CV
                           method = "repeatedcv",
                           number = 10,
                           ## repeated ten times
                           repeats = 5)
```

4.2 Logit 回归

```
set.seed(1)
logit <- train(Terminate ~ Sex + MaritalDesc + Department + PerformanceScore + EngagementSurvey +
               method = "glm",
               trControl = fitControl)
table = logit$results
rownames(table) = NULL
```

```
kable(table, format="latex", booktabs=TRUE, caption = " 在重抽样下 Logit 模型的表现", digits = 3) %>
  kable_styling(latex_options=c("HOLD_position"))
```

表 6: 在重抽样下 Logit 模型的表现

parameter	Accuracy	Kappa	AccuracySD	KappaSD
none	0.653	0.096	0.057	0.138

4.3 线性判别分析（LDA）

```
set.seed(1)
lda <- train(Terminate ~ Sex + MaritalDesc + Department + PerformanceScore + EngagementSurvey + Em
                 method = "lda",
                 trControl = fitControl,
               preProc = c("center", "scale"))
table = lda$results
rownames(table) = NULL
kable(table, format="latex", booktabs=TRUE, caption = " 在重抽样下 LDA 模型的表现", digits = 3) %>%
  kable_styling(latex_options=c("HOLD_position"))
```

表 7: 在重抽样下 LDA 模型的表现

parameter	Accuracy	Kappa	AccuracySD	KappaSD
none	0.656	0.104	0.06	0.142

```
trellis.par.set(caretTheme())
densityplot(lda, pch = "|")
```

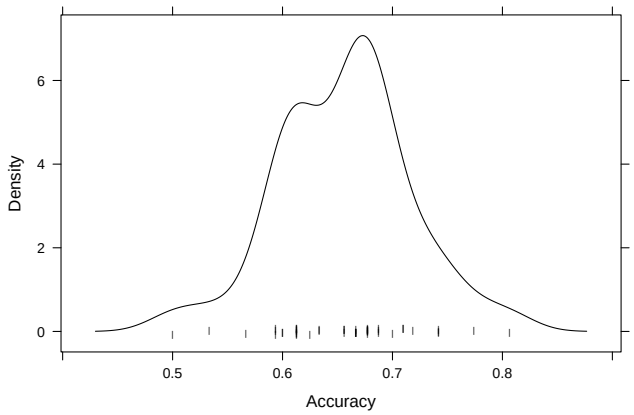

图 7: 在重抽样下 LDA 模型的准确率分布

4.4 偏最小二乘判别分析（PLSDA）

```
set.seed(1)
plsda <- train(Terminate ~ Sex + MaritalDesc + Department + PerformanceScore + EngagementSurvey +
                 method = "pls",
                 trControl = fitControl,
               tuneGrid = expand.grid(.ncomp = 1:10))
table = plsda$results
rownames(table) = NULL
kable(table, format="latex", booktabs=TRUE, caption = " 在重抽样下 PLSDA 模型的表现", digits = 3) %>
  kable_styling(latex_options=c("HOLD_position"))
```

表 8: 在重抽样下 PLSDA 模型的表现

ncomp	Accuracy	Kappa	AccuracySD	KappaSD
1	0.668	0.000	0.010	0.000
2	0.668	0.000	0.010	0.000
3	0.642	-0.018	0.049	0.109
4	0.641	-0.013	0.046	0.105
5	0.642	0.022	0.057	0.136
6	0.649	0.041	0.053	0.128
7	0.671	0.115	0.051	0.127
8	0.673	0.133	0.056	0.139
9	0.668	0.124	0.052	0.125
10	0.657	0.091	0.055	0.137

```r
trellis.par.set(caretTheme())
plot(plsda, metric = "Kappa")
plot(plsda)
```

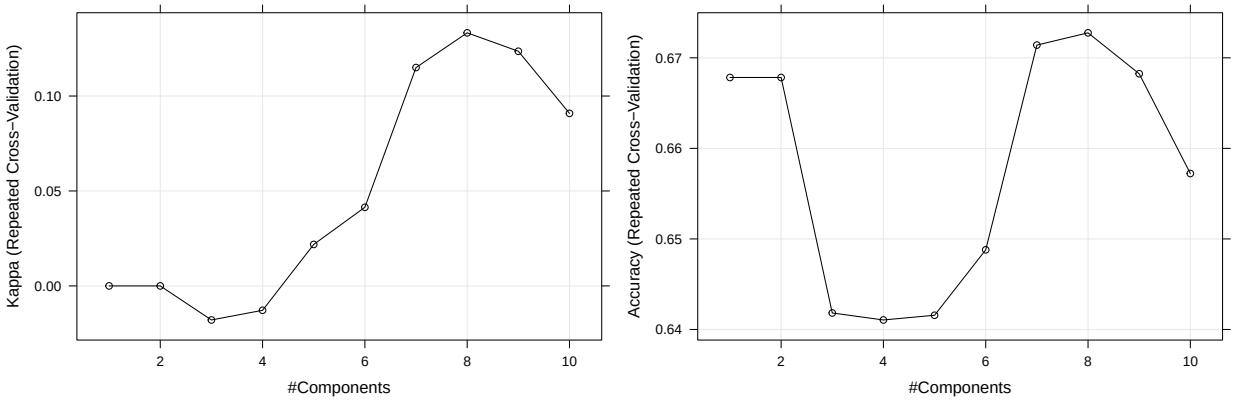

图 8: Kappa 指标和准确率随主成分个数的变化

```r
plsImp = varImp(plsda, scale = FALSE)
table = data.frame(variables = rownames(plsImp$importance), importence = plsImp$importance$Overall
ggplot(table, aes(x = reorder(variables, importence), y = importence)) +
  geom_col() +
  theme_minimal() +
  coord_flip() +
  labs(x = "variables")
```

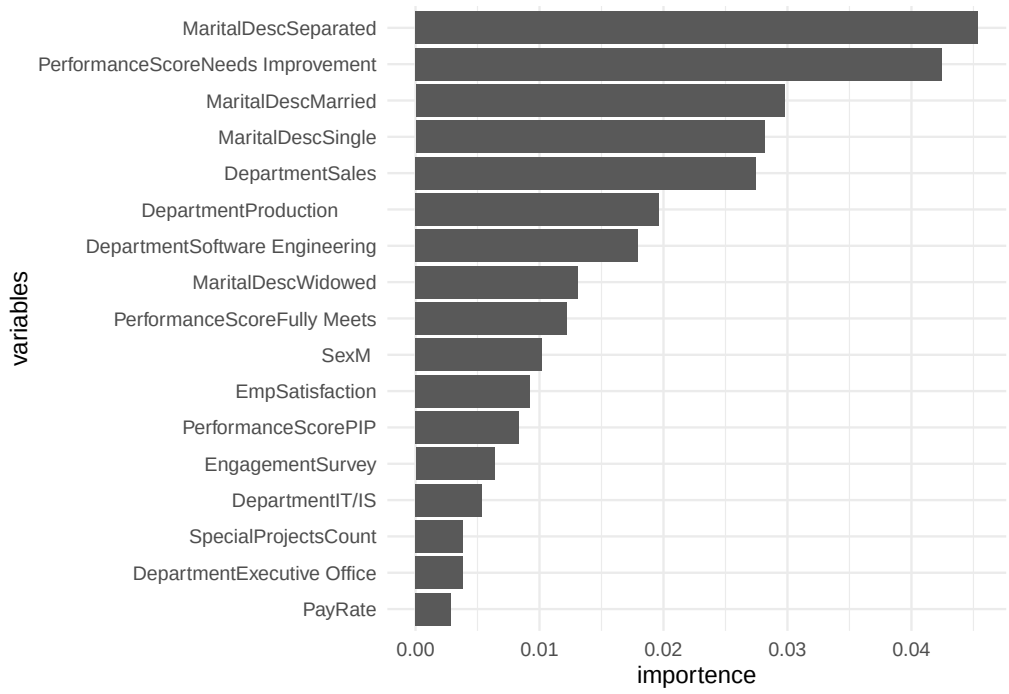

图 9: 变量重要程度

4.5 SVM

```
set.seed(1)
svm <- train(Terminate ~ Sex + MaritalDesc + Department + PerformanceScore + EngagementSurvey + Em
                 method = "svmRadial",
                 trControl = fitControl,
             tuneLength = 5)
kable(svm$results, format="latex", booktabs=TRUE, caption = " 在重抽样下 SVM 模型的表现", digits = 3
  kable_styling(latex_options=c("HOLD_position"))
```

表 9: 在重抽样下 SVM 模型的表现

sigma	C	Accuracy	Kappa	AccuracySD	KappaSD
0.059	0.25	0.668	0.000	0.010	0.000
0.059	0.50	0.662	-0.010	0.021	0.038
0.059	1.00	0.644	0.025	0.053	0.123
0.059	2.00	0.644	0.087	0.069	0.142
0.059	4.00	0.640	0.100	0.071	0.162

```
trellis.par.set(caretTheme())
plot(svm)
plot(svm, metric = "Kappa")
```

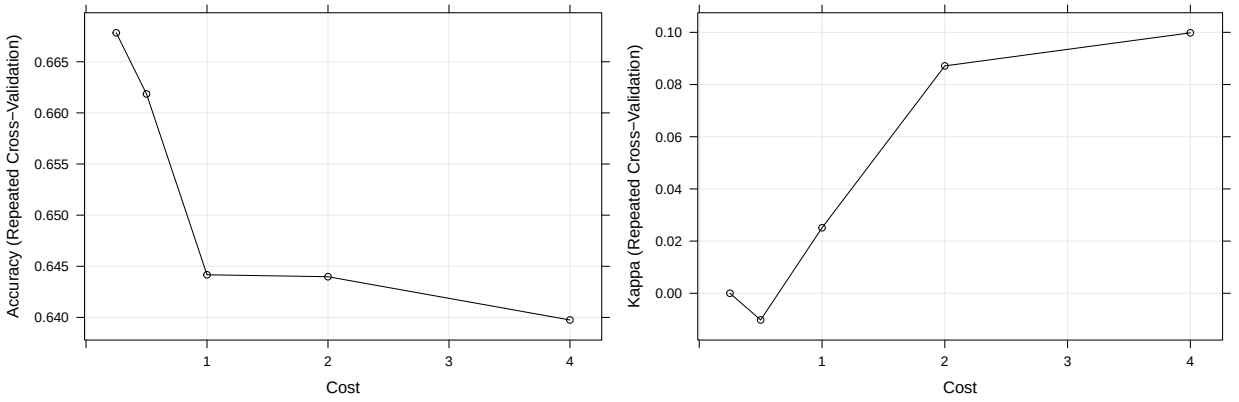

图 10: 调优参数不同取值下的准确率和 Kappa 指标变化

4.6 随机梯度助推法（GBM）

```
trellis.par.set(caretTheme())
plot(gbm)

trellis.par.set(caretTheme())
plot(gbm, metric = "Kappa")
```

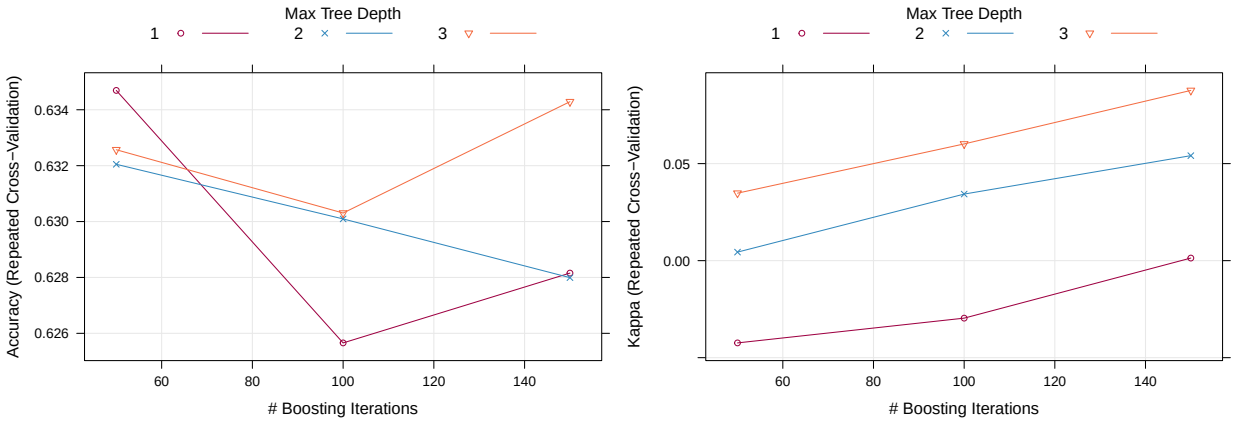

图 11: 调优参数和迭代次数不同取值下的准确率和 Kappa 指标变化

```
trellis.par.set(caretTheme())
densityplot(gbm, pch = "|")
```

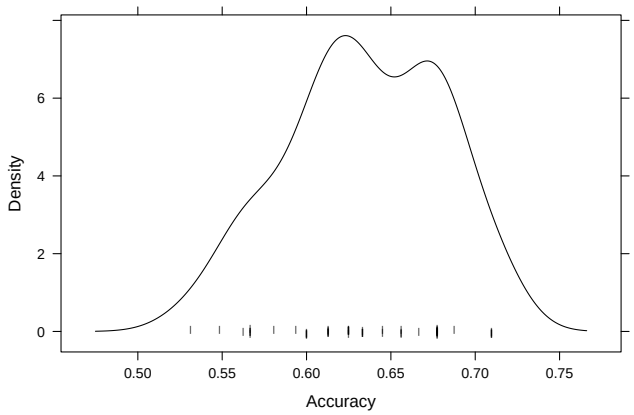

图 12: 在重抽样下 GBM 模型的准确率分布

4.7 模型间的比较

```
resamp = resamples(list(LDA = lda, PLSDA = plsda, SVM = svm, GBM = gbm, Logit = logit))
s1 = summary(resamp)
s2 = summary(diff(resamp))
```

```
ggplot(resamp,
       models = c("LDA", "PLSDA", "GBM", "Logit"),
       metric = "Kappa",
       conf.level = 0.95) +
  theme_bw()
```

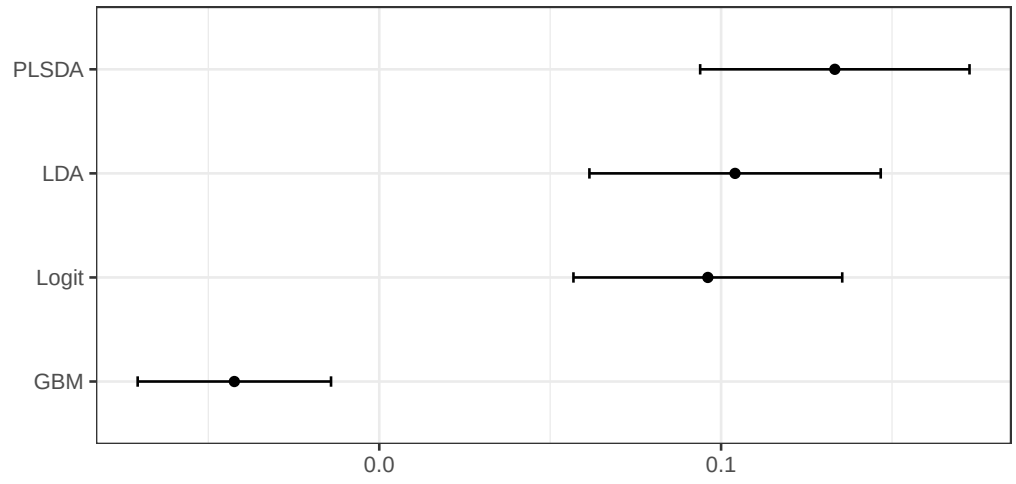

图 13: 模型间 Kappa 的比较（0.95 置信区间）

```
ggplot(resamp,
       models = c("LDA", "PLSDA", "SVM", "GBM", "Logit"),
       metric = "Accuracy",
       conf.level = 0.95) +
  theme_bw()
```

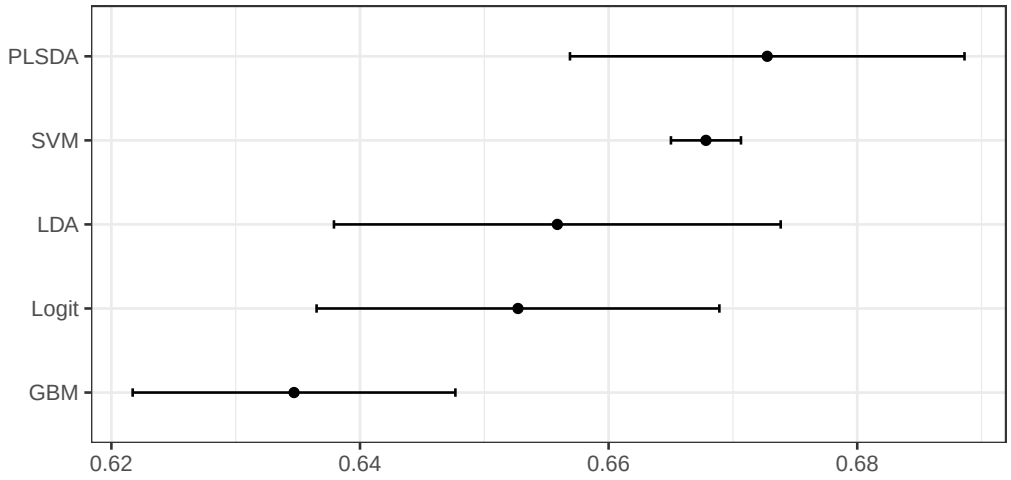

图 14: 模型间准确率的比较（0.95 置信区间）

5 附录

5.1 模型间准确率和 Kappa 的比较

```
kable(s1$statistics$Accuracy, format="latex", booktabs=TRUE, caption = " 模型间准确率的比较", digit
  kable_styling(latex_options=c("HOLD_position"))
```

表 10: 模型间准确率的比较

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
LDA	0.500	0.613	0.667	0.656	0.688	0.806	5
PLSDA	0.533	0.636	0.677	0.673	0.710	0.774	0
SVM	0.645	0.656	0.667	0.668	0.677	0.677	0
GBM	0.531	0.603	0.633	0.635	0.677	0.710	0
Logit	0.533	0.613	0.656	0.653	0.688	0.806	0

```
kable(s2$table$Accuracy, format="latex", booktabs=TRUE, caption = " 模型间准确率差异矩阵", digit = 3
  kable_styling(latex_options=c("HOLD_position"))
```

表 11: 模型间准确率差异矩阵

	LDA	PLSDA	SVM	GBM	Logit
LDA		-0.018026	-0.011831	0.020037	0.001408
PLSDA	0.0051181		0.004930	0.038067	0.020051
SVM	1.0000000	1.0000000		0.033137	0.015121
GBM	0.3387653	0.0001695	1.77e-05		-0.018016
Logit	1.0000000	0.0052567	0.6189921	0.4732932	

```
kable(s1$statistics$Kappa, format="latex", booktabs=TRUE, caption = " 模型间 Kappa 的比较", digit =
  kable_styling(latex_options=c("HOLD_position"))
```

表 12: 模型间 Kappa 的比较

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
LDA	-0.216	0.011	0.108	0.104	0.187	0.475	5
PLSDA	-0.235	0.055	0.118	0.133	0.204	0.405	0
SVM	0.000	0.000	0.000	0.000	0.000	0.000	0
GBM	-0.226	-0.120	-0.061	-0.042	0.000	0.187	0
Logit	-0.167	0.000	0.073	0.096	0.187	0.475	0

```
kable(s2$table$Accuracy, format="latex", booktabs=TRUE, caption = " 模型间 Kappa 差异矩阵", digit =
  kable_styling(latex_options=c("HOLD_position"))
```

表 13: 模型间 Kappa 差异矩阵

	LDA	PLSDA	SVM	GBM	Logit
LDA		-0.018026	-0.011831	0.020037	0.001408
PLSDA	0.0051181		0.004930	0.038067	0.020051
SVM	1.0000000	1.0000000		0.033137	0.015121
GBM	0.3387653	0.0001695	1.77e-05		-0.018016
Logit	1.0000000	0.0052567	0.6189921	0.4732932	

5.2 Logit 回归结果

```
logit2_sum
```

```
##
## Call:
## glm(formula = Terminate ~ Sex + MaritalDesc + Department + PerformanceScore +
##     EngagementSurvey + EmpSatisfaction + SpecialProjectsCount +
##     PayRate, family = binomial(link = "logit"), data = dat_complete)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4867  -0.8979  -0.5757   1.1555   2.2988
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       0.69743    1.51500   0.460  0.64527
## SexM                             -0.04172    0.26676  -0.156  0.87573
## MaritalDescMarried               -0.56882    0.42858  -1.327  0.18444
## MaritalDescSeparated             -2.59543    1.12806  -2.301  0.02140 *
## MaritalDescSingle                -1.15352    0.43198  -2.670  0.00758 **
## MaritalDescWidowed                0.07998    0.85223   0.094  0.92523
## DepartmentIT/IS                   1.32151    1.21236   1.090  0.27570
## DepartmentProduction             -1.05342    1.09945  -0.958  0.33799
## DepartmentSales                  -2.10424    1.22419  -1.719  0.08564 .
## DepartmentSoftware Engineering    1.51244    1.22450   1.235  0.21677
## PerformanceScoreFully Meets       0.77312    0.44795   1.726  0.08436 .
## PerformanceScoreNeeds Improvement 1.60536    0.65051   2.468  0.01359 *
## PerformanceScorePIP               1.13754    0.83610   1.361  0.17366
## EngagementSurvey                  0.02163    0.10257   0.211  0.83296
## EmpSatisfaction                   0.05923    0.15721   0.377  0.70633
## SpecialProjectsCount             -0.52495    0.27532  -1.907  0.05656 .
## PayRate                          -0.01357    0.01503  -0.903  0.36644
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 393.37  on 308  degrees of freedom
```

```
## Residual deviance: 353.68  on 292  degrees of freedom
## AIC: 387.68
##
## Number of Fisher Scoring iterations: 5
```

5.3 数据

```
str(dat)
```

```
## 'data.frame':    310 obs. of  34 variables:
##  $ EmpID                     : int  1103024456 1106026572 1302053333 1211050782 1307059817 7110
##  $ MarriedID                 : int  1 0 0 1 0 1 1 0 0 1 ...
##  $ MaritalStatusID           : int  1 2 0 1 0 1 1 0 0 1 ...
##  $ GenderID                  : int  0 1 1 0 0 0 0 0 0 1 ...
##  $ EmpStatusID               : int  1 1 1 1 1 5 5 1 1 1 ...
##  $ DeptID                    : int  1 1 1 1 1 1 6 6 6 6 ...
##  $ PerfScoreID               : int  3 3 3 3 3 3 3 3 1 3 ...
##  $ FromDiversityJobFairID    : int  1 0 0 0 0 1 0 0 0 0 ...
##  $ PayRate                   : num  28.5 23 29 21.5 16.6 ...
##  $ Termd                     : int  0 0 0 1 0 1 1 0 0 0 ...
##  $ PositionID                : int  1 1 1 2 2 2 3 3 3 3 ...
##  $ Position                  : Factor w/ 32 levels "Accountant I",..: 1 1 1 2 2 2 3 3 3 3 ...
##  $ State                     : Factor w/ 28 levels "AL","AZ","CA",..: 11 11 11 11 11 11 26 27 2
##  $ Zip                       : int  1450 1460 2703 2170 2330 1844 21851 5664 98052 3062 ...
##  $ DOB                       : Factor w/ 306 levels "01/02/51","01/04/64",..: 283 87 204 223 11
##  $ Sex                       : Factor w/ 2 levels "F","M ": 1 2 2 1 1 1 1 1 1 2 ...
##  $ MaritalDesc               : Factor w/ 5 levels "Divorced","Married",..: 2 1 4 2 4 2 2 4 4 2
##  $ CitizenDesc               : Factor w/ 3 levels "Eligible NonCitizen",..: 3 3 3 3 3 3 1 3 3 3
##  $ HispanicLatino            : Factor w/ 4 levels "no","No","yes",..: 2 2 2 2 2 2 2 2 4 2 ...
##  $ RaceDesc                  : Factor w/ 6 levels "American Indian or Alaska Native",..: 3 3 6
##  $ DateofHire                : Factor w/ 99 levels "1/10/2011","1/20/2013",..: 20 9 94 32 52 92
##  $ DateofTermination         : Factor w/ 94 levels "","01/15/16",..: 1 1 1 13 1 41 81 1 1 1 ...
##  $ TermReason                : Factor w/ 18 levels "","Another position",..: 12 12 12 1 12 4 2
##  $ EmploymentStatus          : Factor w/ 5 levels "Active","Future Start",..: 1 1 1 4 1 5 5 1 1
##  $ Department                : Factor w/ 6 levels "Admin Offices",..: 1 1 1 1 1 1 5 5 5 5 ...
##  $ ManagerName               : Factor w/ 21 levels "Alex Sweetwater",..: 4 4 4 4 4 4 13 13 13 1
##  $ ManagerID                 : int  1 1 1 1 1 1 17 17 17 17 ...
##  $ RecruitmentSource         : Factor w/ 23 levels "Billboard","Careerbuilder",..: 4 22 9 17 22
##  $ PerformanceScore          : Factor w/ 4 levels "Exceeds","Fully Meets",..: 2 2 2 2 2 2 2 2 4
```

```
##  $ EngagementSurvey          : num  2.04 5 3.9 3.24 5 3.8 3.14 5 2.3 3.6 ...
##  $ EmpSatisfaction           : int  2 4 5 3 3 4 5 5 1 5 ...
##  $ SpecialProjectsCount      : int  6 4 5 4 5 4 0 0 0 0 ...
##  $ LastPerformanceReview_Date: Factor w/ 43 levels "","1/10/2019",..: 5 7 8 1 5 1 1 11 18 20 ..
##  $ DaysLateLast30            : int  0 0 0 NA 0 NA NA 0 0 0 ...
```

```
summary(dat)
```

```
##      EmpID              MarriedID      MaritalStatusID     GenderID
##  Min.   :6.020e+08   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:1.101e+09   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :1.203e+09   Median :0.0000   Median :1.0000   Median :0.000
##  Mean   :1.200e+09   Mean   :0.3968   Mean   :0.8097   Mean   :0.429
##  3rd Qu.:1.379e+09   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.000
##  Max.   :1.988e+09   Max.   :1.0000   Max.   :4.0000   Max.   :1.000
##
##   EmpStatusID        DeptID       PerfScoreID    FromDiversityJobFairID
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :0.00000
##  1st Qu.:1.000   1st Qu.:5.000   1st Qu.:3.000   1st Qu.:0.00000
##  Median :1.000   Median :5.000   Median :3.000   Median :0.00000
##  Mean   :2.397   Mean   :4.606   Mean   :2.984   Mean   :0.09355
##  3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:3.000   3rd Qu.:0.00000
##  Max.   :5.000   Max.   :6.000   Max.   :4.000   Max.   :1.00000
##
##     PayRate          Termd          PositionID
##  Min.   :14.00   Min.   :0.0000   Min.   : 1.00
##  1st Qu.:20.00   1st Qu.:0.0000   1st Qu.:18.00
##  Median :24.00   Median :0.0000   Median :19.00
##  Mean   :31.28   Mean   :0.3323   Mean   :16.84
##  3rd Qu.:45.31   3rd Qu.:1.0000   3rd Qu.:20.00
##  Max.   :80.00   Max.   :1.0000   Max.   :30.00
##
##                      Position       State          Zip              DOB
##  Production Technician I :136   MA     :275   Min.   : 1013   06/14/87:  2
##  Production Technician II: 57   CT     :  6   1st Qu.: 1901   07/07/84:  2
##  Area Sales Manager      : 27   TX     :  3   Median : 2132   09/09/65:  2
##  Production Manager      : 14   VT     :  2   Mean   : 6570   09/22/76:  2
##  Software Engineer       :  9   AL     :  1   3rd Qu.: 2357   01/02/51:  1
##  IT Support              :  8   AZ     :  1   Max.   :98052   01/04/64:  1
##  (Other)                 : 59   (Other): 22                   (Other) :300
```

```
##  Sex         MaritalDesc               CitizenDesc  HispanicLatino
##  F :177   Divorced : 30   Eligible NonCitizen: 12   no :  1
##  M :133   Married  :123   Non-Citizen        :  4   No :281
##           Separated: 12   US Citizen         :294   yes:  1
##           Single   :137                              Yes: 27
##           Widowed  :  8
##
##
##                                RaceDesc       DateofHire  DateofTermination
##  American Indian or Alaska Native:  4   1/10/2011: 14           :207
##  Asian                            : 34   3/30/2015: 12   08/19/13 :  2
##  Black or African American        : 57   1/5/2015 : 11   09/24/12 :  2
##  Hispanic                         :  4   9/29/2014: 11   09/26/11 :  2
##  Two or more races                : 18   5/16/2011: 10   2001/9/12:  2
##  White                            :193   7/5/2011 : 10   2004/1/13:  2
##                                          (Other)  :242   (Other)  : 93
##                      TermReason               EmploymentStatus
##  N/A - still employed     :196   Active                :182
##  Another position         : 20   Future Start          : 11
##  unhappy                  : 14   Leave of Absence      : 14
##  more money               : 11   Terminated for Cause  : 15
##  N/A - Has not started yet: 11   Voluntarily Terminated: 88
##  career change            :  9
##  (Other)                  : 49
##                 Department          ManagerName    ManagerID
##  Admin Offices       : 10   Elijiah Gray  : 22   Min.   : 1.00
##  Executive Office    :  1   Kelley Spirea : 22   1st Qu.:10.00
##  IT/IS               : 50   Kissy Sullivan: 22   Median :15.00
##  Production          :208   Michael Albert: 22   Mean   :14.58
##  Sales               : 31   Amy Dunn      : 21   3rd Qu.:19.00
##  Software Engineering: 10   Brannon Miller: 21   Max.   :39.00
##                             (Other)       :180   NA's   :8
##                         RecruitmentSource          PerformanceScore
##  Employee Referral                : 31     Exceeds          : 37
##  Diversity Job Fair               : 29     Fully Meets      :243
##  Search Engine - Google Bing Yahoo: 25     Needs Improvement: 18
##  Monster.com                      : 24     PIP              : 12
##  Pay Per Click - Google           : 21
##  Professional Society             : 20
```

```
##  (Other)                          :160
##  EngagementSurvey EmpSatisfaction SpecialProjectsCount
##  Min.   :1.030    Min.   :1.00    Min.   :0.00
##  1st Qu.:2.083    1st Qu.:3.00    1st Qu.:0.00
##  Median :3.470    Median :4.00    Median :0.00
##  Mean   :3.332    Mean   :3.89    Mean   :1.21
##  3rd Qu.:4.520    3rd Qu.:5.00    3rd Qu.:0.00
##  Max.   :5.000    Max.   :5.00    Max.   :8.00
##
##  LastPerformanceReview_Date DaysLateLast30
##           :103              Min.   :0
##  1/14/2019: 18              1st Qu.:0
##  2/18/2019: 12              Median :0
##  1/21/2019: 10              Mean   :0
##  1/28/2019:  9              3rd Qu.:0
##  2/25/2019:  9              Max.   :0
##  (Other)  :149              NA's   :103
```