Inadimplência - Modelo Regressão

February 8, 2022

Diego de Almeida Miranda - 133603

1 Introdução

A inadimplência no setor bancário é uma das variáveis mais observadas na área. Para realizar o cálculo da inadimplência, já foram propostos diversos métodos diferentes. Há uma dificuldade em definir precisamente o que é a inadimplência, o que leva os pesquisadores a deixarem de medir diversos desses casos (de inadimplência). Objetivamente, podemos definir a inadimplência como a falta do cumprimento de uma obrigação. Entretanto, apesar de possuir uma definição clara, a tarefa de definir um inadimplente para analisar o risco de crédito não é trivial. Por conta disso, diversos autores apresentam e usam diferentes definições para inadimplência, mas todas tendo em foco um devedor que não sanou suas dívidas.

Ainda assim, devido ao mercado e nível de competitividade empresarial crescente no decorrer dos anos, se faz necessária maior eficiência no gerenciamento dos recursos capitais de todas as organizações. O acesso facilitado ao crédito também abriu portas para a concorrência de maus pagadores. Nitidamente, a existência de mau pagadores é um problema para as empresas de todo porte, pois na concessão de crédito é feito um planejamento de quando obteria novamente aquele valor.Ainda que as empresas e consumidores tenham aproximado relações nos últimos anos, através de diversas obrigações contratuais, é corriqueiro que haja o descumprimento destes acordos. Isso tanto de quebra de acordo do cliente para a empresa, quanto da quebra da empresa para com o cliente.

Com um eventual não cumprimento do cliente com as cláusulas estabelecidas no contrato, como o não pagamento do valor acordado, a empresa precisa tomar medidas, podendo cancelar o serviço que está sendo prestado, até que o cliente acerte suas dívidas com a empresa em questão. Mesmo assim, é necessário que a empresa tenha maneiras de se proteger contra clientes com alto potencial de se tornarem inadimplentes.

O Serasa realizou um levantamento em 2021 indicando que em Abril haviam 62,98 milhões de brasileiros que estão inadimplentes. Em setembro, esta dívida atingiu o valor de $R\$\$245,3 Bilhes. Nomesmoms, havia R\$\$3.944,65$ de dívida por pessoa inadimplente. Além disso, bancos e cartões de crédito são os credores de 28,70% das dívidas, sendo seguidos pelas empresas relacionadas ao abastecimento de água, luz e contas básicas, correspondendo a 23,5% das dívidas. Empresas do comércio varejistas aparecem ocupando o terceiro lugar, com 13% das dívidas totais no Brasil.

Para fins de estudo, utilizaremos este conjunto de dados sintéticos obtidos através da plataforma Kaggle, intitulado como *Loan Default Prediction*. Segundo o autor, são dados obtidos através

de instituições financeiras, mas que por segurança das instituições envolvida, foram removidos quaisquer dados que possam ser utilizados para rastrear e identificar algum desses clientes.

Aqui, temos 3 principais atributos para identificar uma amostra de indivíduo/instituição inadimplente, sendo elas a condição de emprego, se está empregado ou não, qual o valor de dinheiro no banco no momento atual, e por fim qual o seu salário anual. Através destas variáveis teremos de realizar as análises que poderão dizer ou não quando um indivíduo tem chances consideráveis de ser inadimplente. Normalmente, para este tipo de análise, também são considerados os dados de percepção do mercado, setor de atividade da empresa/individuo, tempo de atividade, nível de informatização da empresa, nível de escolaridade do sócio, entre outros. Dependendo da população estudada, i.e., microempresa, empresas médias, de grande porte ou até mesmo indivíduos, esses atributos podem variar.

É notável que o volume de dados é algo a ser fortemente considerado, uma vez que não podemos fazer as análises puramente através de análises gráficas e exploratória dos dados. A grande quantidade de amostras e parâmetros para serem analisados torna a tarefa de compreender os padrões de comportamento dos clientes uma atividade extremamente complicada, mesmo que para grandes equipes de pessoas.

Em advento do desenvolvimento tecnológico e matemático, podemos aplicar métodos de aprendizado de máquina para encontrar relações entre as variáveis descritivas e objetivas do problema. Quando pensamos em dados de inadimplência, podemos buscar compreender quanto a relação entre o salário anual e o extrato do banco de um indivíduo interfere no problema de classificação de um indivíduo inadimplente.

Para este conjunto de dados em específico, foram escolhidos métodos de classificação a partir da regressão logística, modelo este mais indicado pela literatura sobre o tema. O fato de termos uma variável objetivo binária, ou é inadimplente ou não é, torna o uso da regressão linear convencional algo impróprio, uma vez que não estamos prevendo dados contínuos. A adequação de um modelo para com os dados pode ser feita através das análises de significância estatística das variáveis descritivas. Além da Regressão logística convencional, será utilizado a regressão logística com penalização e, por fim, o método de classificação Random Forest.

Como estamos trabalhando com variáveis categóricas, não existem razões para utilizar um método de regressão que nos retorne valores contínuos. Da mesma forma, vale para a métrica de avaliação que será aqui utilizada. Os modelos aqui apresentados serão avaliados utilizando as métricas de acurácia, área sob a ROC - conhecido como AUC ROC - e pontuação F1. A acurácia é dada pela divisão entre a soma dos resultados verdadeiros (positivo e negativo) pela quantidade de amostras que temos. Sendo VP amostras classificadas corretamente como positivas, e VN amostras classificadas corretamente como negativas, então a acurácia acc pode ser dada por $acc = \frac{VP+VN}{|amostras|}$. A métrica de pontuação F1 é a média harmônica entre precisão e revocação, e é dada pela seguinte equação $f1 = 2\frac{Precisao \times Revocaao}{Precisao+Revocaao}$. Por fim, a curva ROC é realizada através das taxas de verdadeiro positivo e verdadeiro negativo. No eixo X da curva ROC temos o *False positive Rate*, onde $FPR = \frac{FP}{VP+FP}$. De forma análoga, temos no eixo Y temos o *True Positive Rate*, dado por $FPR = \frac{VP}{VP+FN}$. A métrica de AUC nada mais é do que o valor da área sob a curva ROC.

Nenhuma métrica de validação cruzada será implementada neste trabalho, o que pode acabar por não assegurar a capacidade de generalização dos modelos. Por fim, podemos dizer que todas as análises aqui apresentadas foram feitas utilizando a linguagem de programação R e suas bibliotecas.

2 Análise Exploratória

Importando todas as bibliotecas necessárias e potencialmente uteis para o estudo.

[ ]:
```
# instalando e importando as bibliotecas necessárias
install.packages('pacman')
base::library(pacman)
pacman::p_load(car, nortest, psych, olsrr, corrplot, ggplot2, MASS, caret,␣
↪tidyverse, graphics,stats,splimes,pROC, dplyr, broom)
```

[ ]:
```
library('MLmetrics')
```

[ ]:
```
library(randomForest)
```

[ ]:
```
install.packages('randomForest')
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

[ ]:
```
# leitura do conjunto de dados e remoção da variáveis indice
df <- utils::read.csv("./Default_Fin.csv")
df<- df[ , !(names(df) %in% c('Index'))]
head(df)
```

A data.frame: 6 × 4

	Employed <int>	Bank.Balance <dbl>	Annual.Salary <dbl>	Defaulted. <int>
1	1	8754.36	532339.56	0
2	0	9806.16	145273.56	0
3	1	12882.60	381205.68	0
4	1	6351.00	428453.88	0
5	1	9427.92	461562.00	0
6	0	11035.08	89898.72	0

[ ]:
```
#conferindo se existem valores na no conjunto de dados
which(is.na(df))
```

[ ]:
```
summary((df))
```

```
   Employed       Bank.Balance   Annual.Salary      Defaulted.
Min.   :0.0000   Min.   :    0   Min.   :  9264   Min.   :0.0000
1st Qu.:0.0000   1st Qu.: 5781   1st Qu.:256086   1st Qu.:0.0000
Median :1.0000   Median : 9884   Median :414632   Median :0.0000
Mean   :0.7056   Mean   :10024   Mean   :402204   Mean   :0.0333
3rd Qu.:1.0000   3rd Qu.:13996   3rd Qu.:525693   3rd Qu.:0.0000
Max.   :1.0000   Max.   :31852   Max.   :882651   Max.   :1.0000
```

Até então, pudemos ver que não existem dados faltantes ou corrompidos em nenhuma das amostras. Além disso, nos é apresentado que a média do *Bank Balance* é de 10.024 unidades de medida, e a

média do *Annual Salary* é de 402.204 unidades de medida.

Vejamos agora, quais e quantos são os inadimplentes neste conjunto de dados.

[ ]:
```
# número de inadimplentes no conjunto de dados
nrow(df[sample( which( df$Defaulted. == 1)) , ])
```

333

[ ]:
```
# porcentagem de inadimplentes
nrow(df[sample( which( df$Defaulted. == 1)) , ])/nrow(df)
```

0.0333

[ ]:
```
#visualizando algumas amostras de inadimplentes
head(df[sample( which( df$Defaulted. == 1)) , ])
```

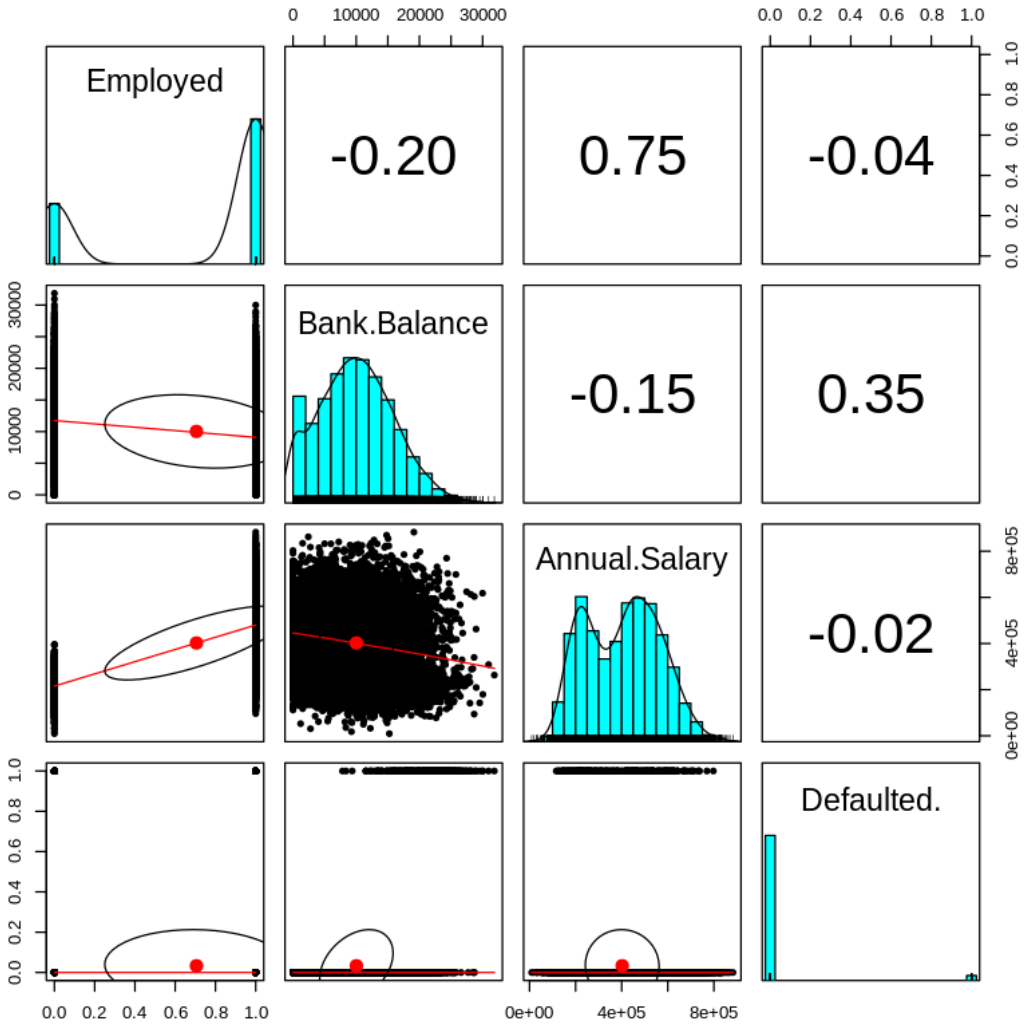
A data.frame: 6 × 4

	Employed <int>	Bank.Balance <dbl>	Annual.Salary <dbl>	Defaulted. <int>
6462	1	25496.04	534240.0	1
3882	1	16107.12	428299.7	1
4710	1	24902.28	490589.9	1
3118	0	15471.00	163494.6	1
5821	1	15105.12	315976.4	1
5190	0	25363.56	253208.6	1

Portanto, temos um conjunto com 10.000 amostras, onde 3.33% delas são compostas por pessoas inadimplentes, isto é, 333 amostras de pessoas inadimplentes.

Abaixo podemos ver uma representação gráfica que sintetiza bem diversos aspectos deste conjunto de dados. Não só a distribuição de cada um dos atributos (na diagonal principal), mas como também os indices de correlação entre cada variável(no triangulo superior) e a representação gráfica desta relação (triangulo inferior).

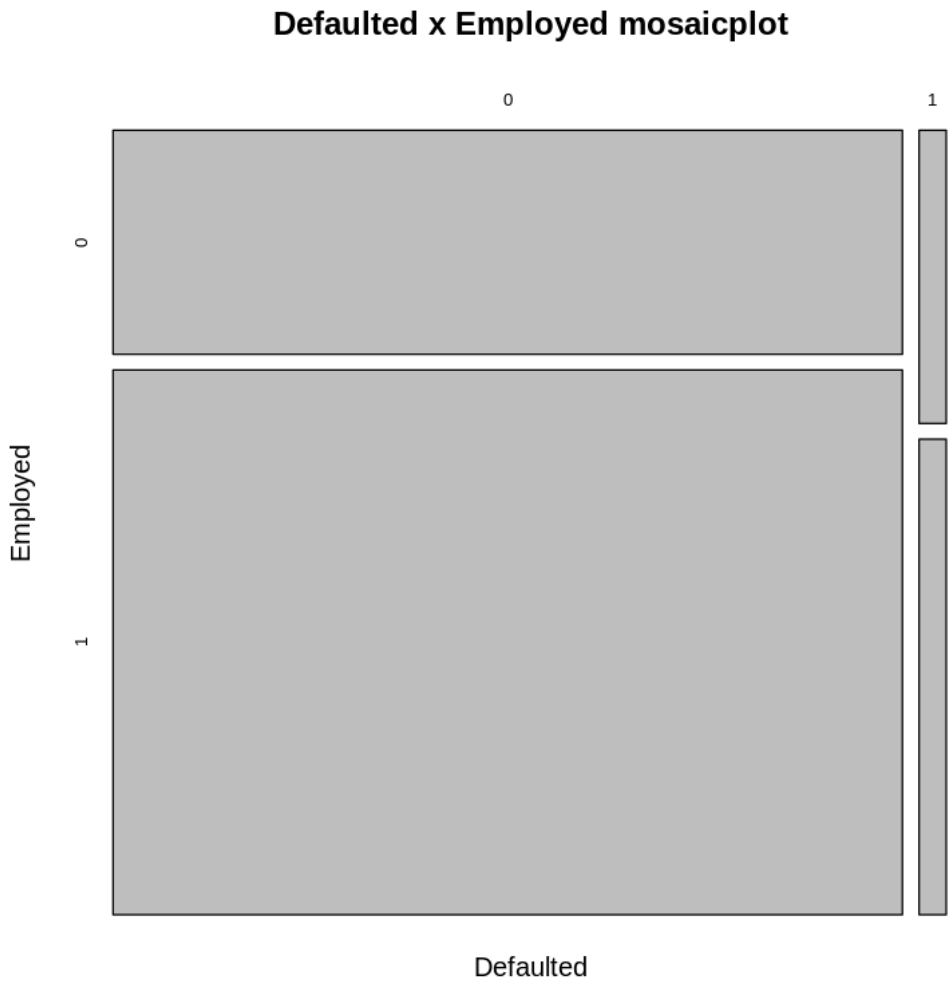
[ ]:
```
psych::pairs.panels (df)
```



Aqui fica evidente a forte correlação entre o salário anual e a condição de emprego dos indivíduos, mas de fato isso é algo a ser esperado. No modelo, é bom que evitemos manter variáveis descritivas com alta correlação, pois isso pode prejudicar o desempenho preditivo e computacional do nosso modelo. Ainda assim, serão mantidas todas as variáveis para um modelo inicial. A variável *Bank Balance* mostra maior correlação com a variável objetivo *Defaulted*.

Vejamos agora de maneira gráfica, como as duas únicas variáveis categóricas no nosso conjunto de dados se relacionam.

[ ]:
```
mosaicplot(table(df$Defaulted., df$Employed), xlab = 'Defaulted',␣
↪ylab='Employed', main='Defaulted x Employed mosaicplot')
```



Dessa forma, a maior parte do conjunto de dados é composto por pessoas não inadimplentes e empregadas, enquanto a menor parte dos dados são formados por pessoas desempregadas e inadimplentes. Mas qual a porcentagem de pessoas inadimplentes e desempregadas? E qual a porcentagem de pessoas inadimplentes e empregadas?

[ ]:
```
# Porcentagem de Empregados e inadimplentes em relação às amostras empregadas
nrow(df[ sample( which( df$Employed == 1 & df$Defaulted. == 1 )) , ])/nrow(df[
↪sample( which( df$Employed == 1)) , ])
```

0.0291950113378685

[ ]:
```
# Porcentagem de desempregados e inadimplentes em relação às amostras de
↪desempregadas
```

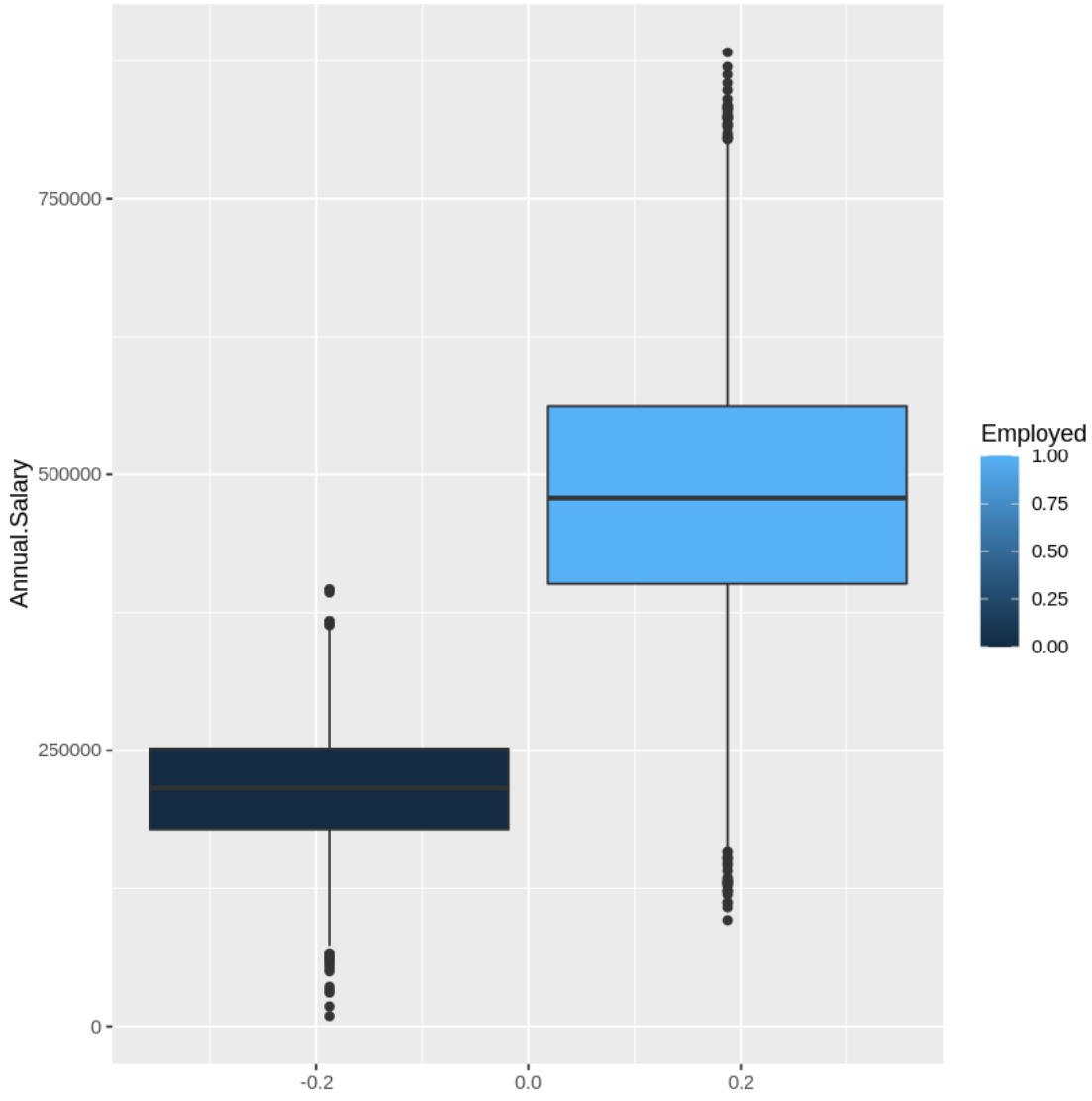
```
nrow(df[ sample( which( df$Employed == 0 & df$Defaulted. == 1 )) , ])/nrow(df[␣
↪sample( which( df$Employed == 0)) , ])
```

0.0431385869565217

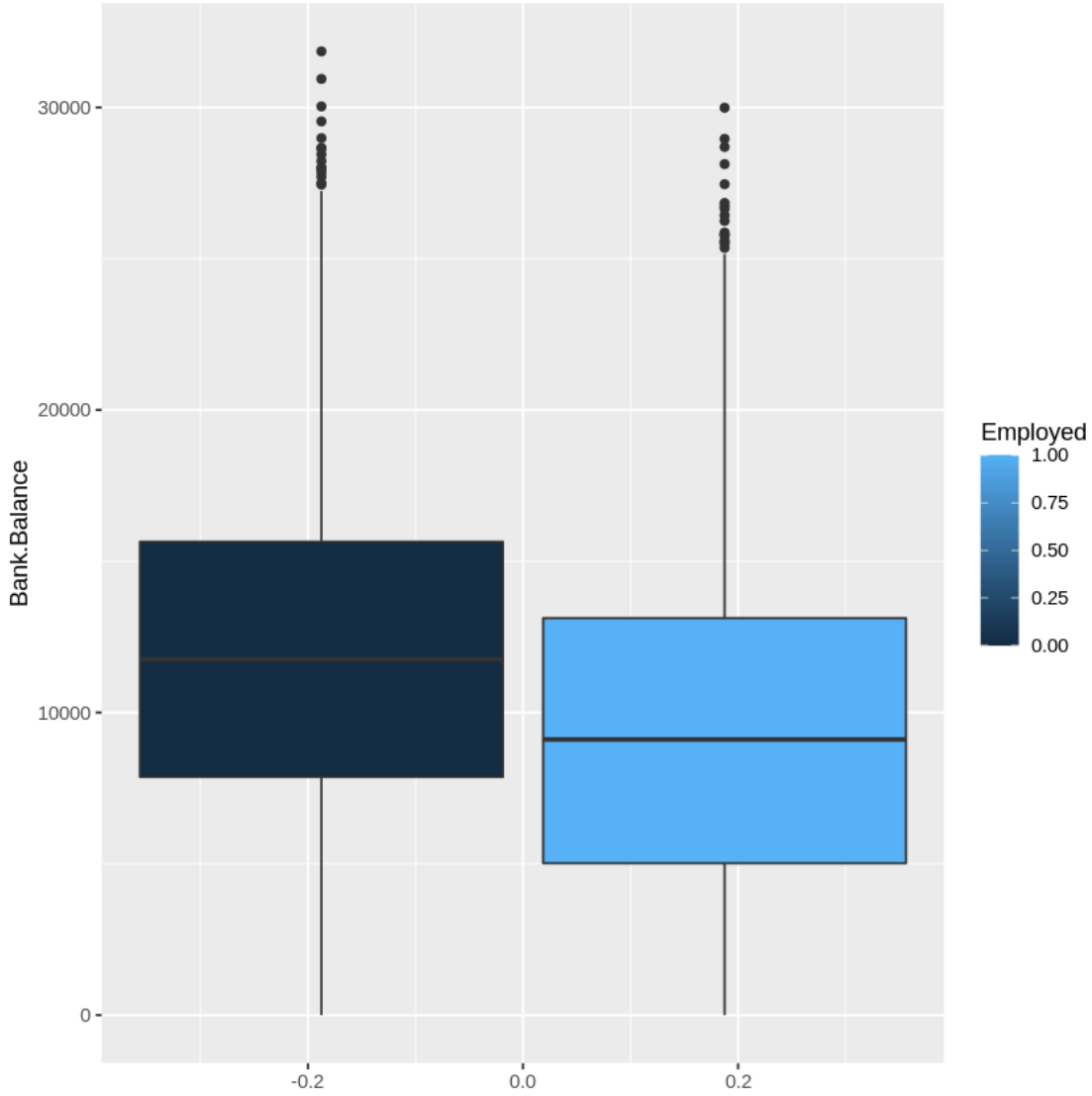
Notamos, não só visualmente, como numéricamente, que a distribuição de indivíduos inadimplentes é similar entre os empregados e não empregados, assim como o indice de correlação de -0.04 sugere.

Vejamos como as variáveis contínuas se relacionam, duas a duas, com as variáveis discretas.

[ ]:
```
# box plot de Salário anual divido entre os empregados e não empregados
ggplot(data = df, aes(y=Annual.Salary)) + geom_boxplot(aes(group=Employed,␣
↪fill=Employed))
```

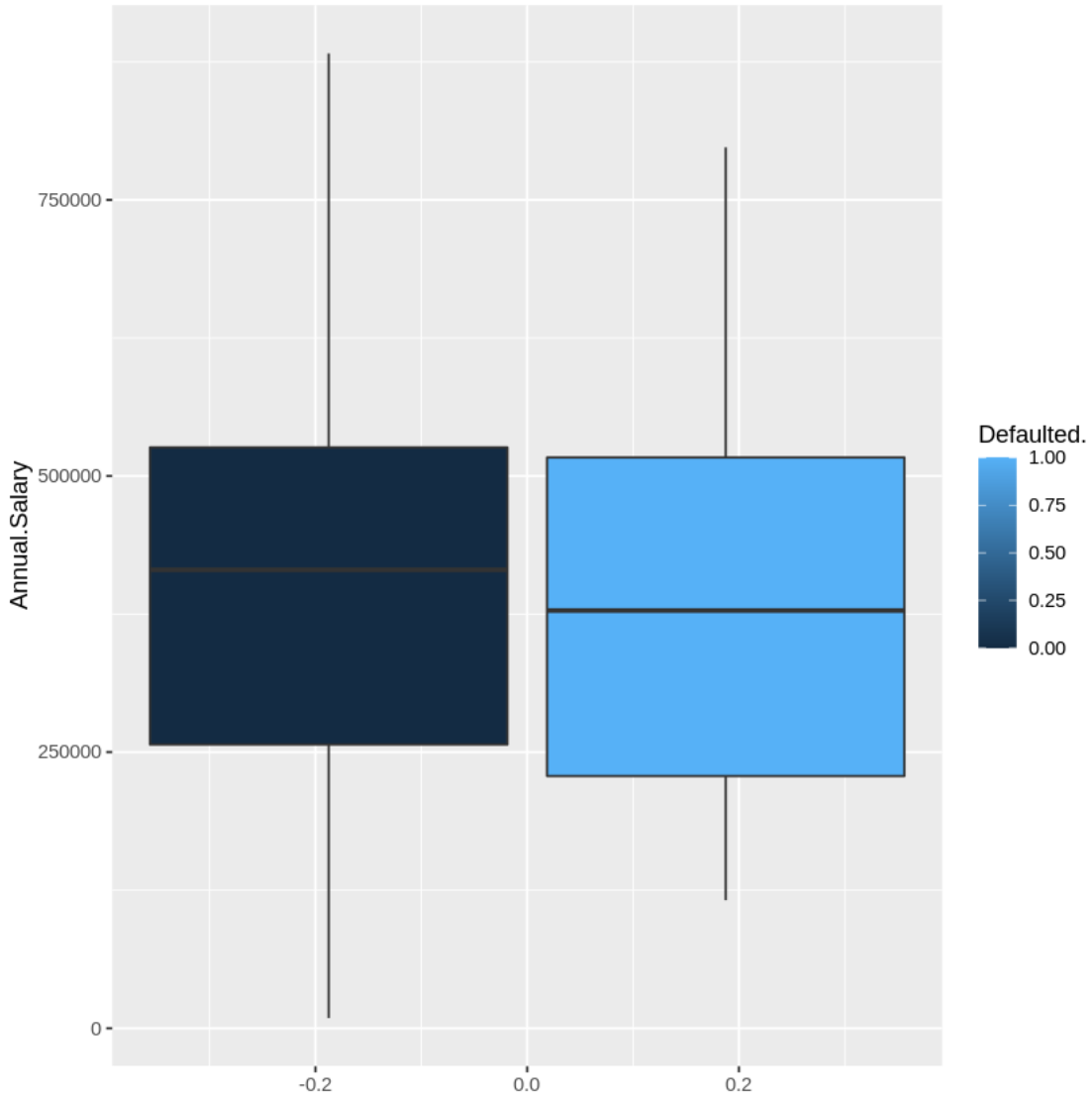


[ ]:
```
# box plot de Bank Balance divido entre os empregados e não empregados
ggplot(data = df, aes(y=Bank.Balance)) + geom_boxplot(aes(group=Employed,␣
↪fill=Employed), show.legend = TRUE)
```

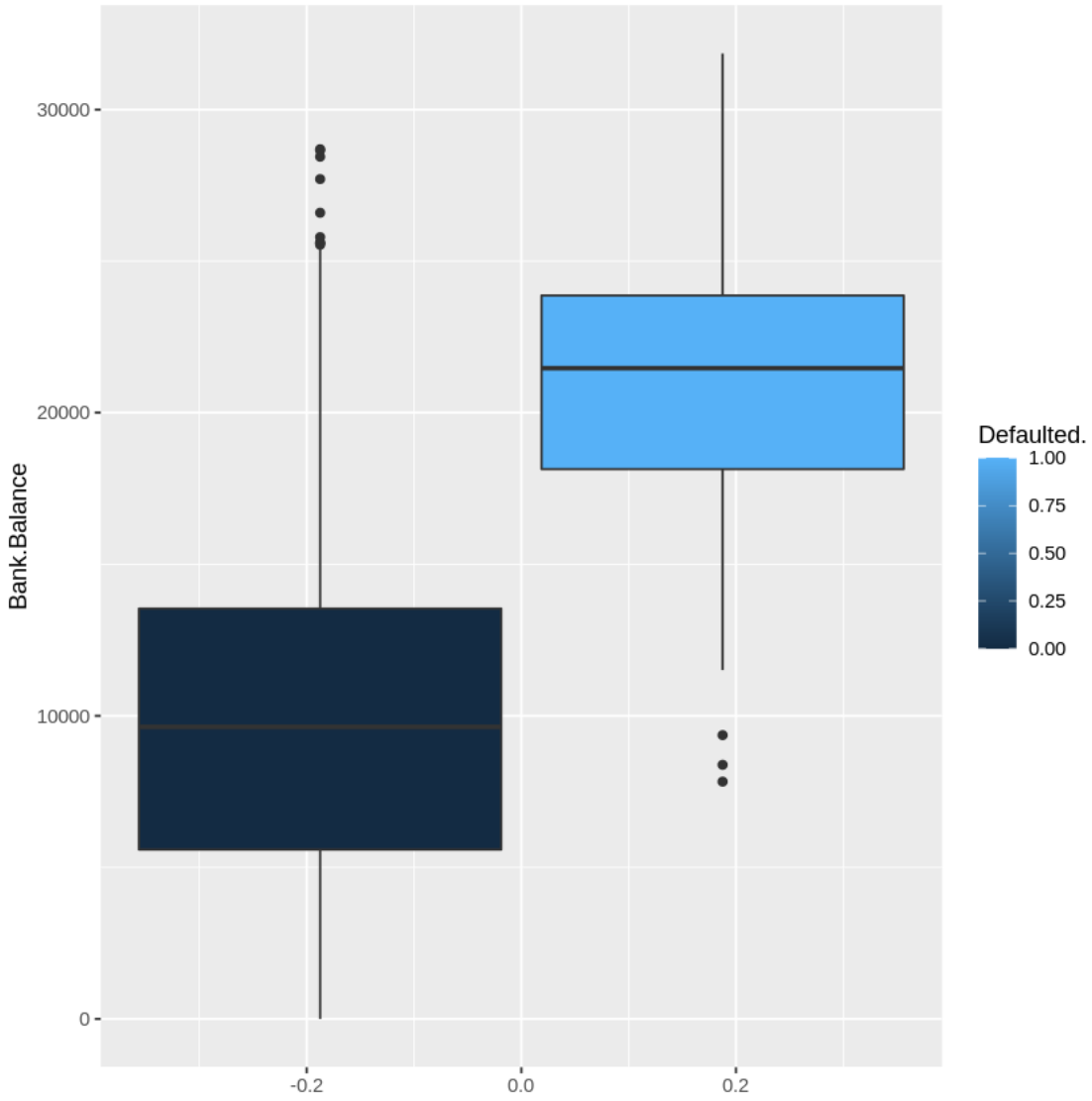


Podemos ver que, normalmente aqueles que estão empregados costumam ter maior salário anual, entretando, a média do valor no banco é maior para os desempregados.

[ ]:
```
# box plot de Salário anual divido entre os inadimplentes e não inadimplentes
ggplot(data = df, aes(y=Annual.Salary)) + geom_boxplot(aes(group=Defaulted.,␣
↪fill=Defaulted.), show.legend = TRUE)
```



[ ]:
```
# box plot de Bank Balance divido entre os inadimplentes e não inadimplentes
ggplot(data = df, aes(y=Bank.Balance)) + geom_boxplot(aes(group=Defaulted.,␣
↪fill=Defaulted.), show.legend = TRUE)
```



Entre os inadimplentes, é notável a pouca relação entre o salário anual, mas o *bank balance* se mostra mais significativo. A média do saldo do banco entre pessoas inadimplentes costuma ser maior que a dos bons pagadores.

3 Construção do Modelo

Inicialmente, devemos separar os conjuntos de dados para treino e para teste. Aqui iremos separar 80% do conjunto de dados para treinar os modelos que iremos utilizar.

[ ]:
```
# aqui estamos defindo um conjunto de dados de 80% das amostras para treino e␣
↪20% para teste.
set.seed(123)
training.samples <- df$Defaulted. %>% createDataPartition(p = 0.8, list = FALSE)
```

```
train.data  <- df[training.samples, ]
test.data <- df[-training.samples, ]
```

3.1 Regressão Logística

Abaixo iremos definir o modelo com todas as variáveis descritivas disponíveis no conjunto de dados.

[ ]:
```
model_logit <- glm( Defaulted. ~ Employed + Bank.Balance + Annual.Salary,
                 data = train.data, family = binomial)
```

Uma vez definido o modelo, podemos avaliar como ele se ajustou aos dados. Vejamos agora um resumo dos ajustes.

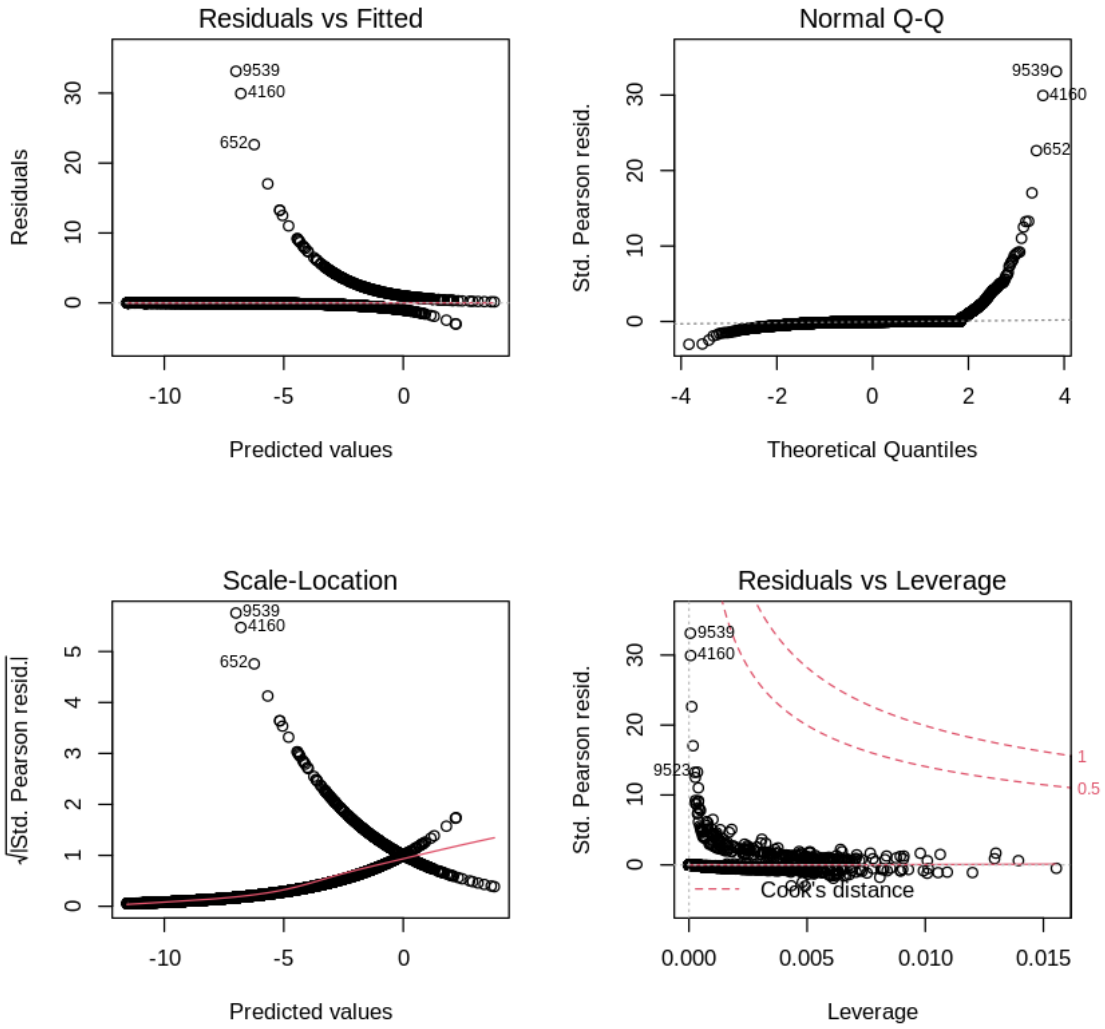
[ ]:
```
summary(model_logit)
```

```
Call:
glm(formula = Defaulted. ~ Employed + Bank.Balance + Annual.Salary,
    family = binomial, data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1525  -0.1404  -0.0559  -0.0199   3.7420

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.161e+01  4.923e-01 -23.572   <2e-16 ***
Employed       6.338e-01  2.621e-01   2.418   0.0156 *
Bank.Balance   4.809e-04  2.171e-05  22.158   <2e-16 ***
Annual.Salary  3.714e-07  7.579e-07   0.490   0.6241
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2340.2  on 7994  degrees of freedom
Residual deviance: 1252.2  on 7991  degrees of freedom
AIC: 1260.2

Number of Fisher Scoring iterations: 8
```

Ao vermos os coeficientes, podemos notar que o mais significativo dos atributos foi o *Bank Balance*, enquanto o *Annual Salary* não é uma variável significatia nesse modelo. Obtvemos também um valor no critério de informação Akaike (AIC) de 1260.3. Neste critério, bons modelos são aqueles com menor valor. Outra informação aqui disponível é que nosso modelo com todos os parâmetros presentes tem mais assertividade, ou se ajusta melhor aos dados, do que um modelo nulo (apenas com o intercepto).

Vejamos agora alguns plots para ajudar na análise gráfica.

[ ]:
```
par(mfrow=c(2,2))
plot(model_logit)
```



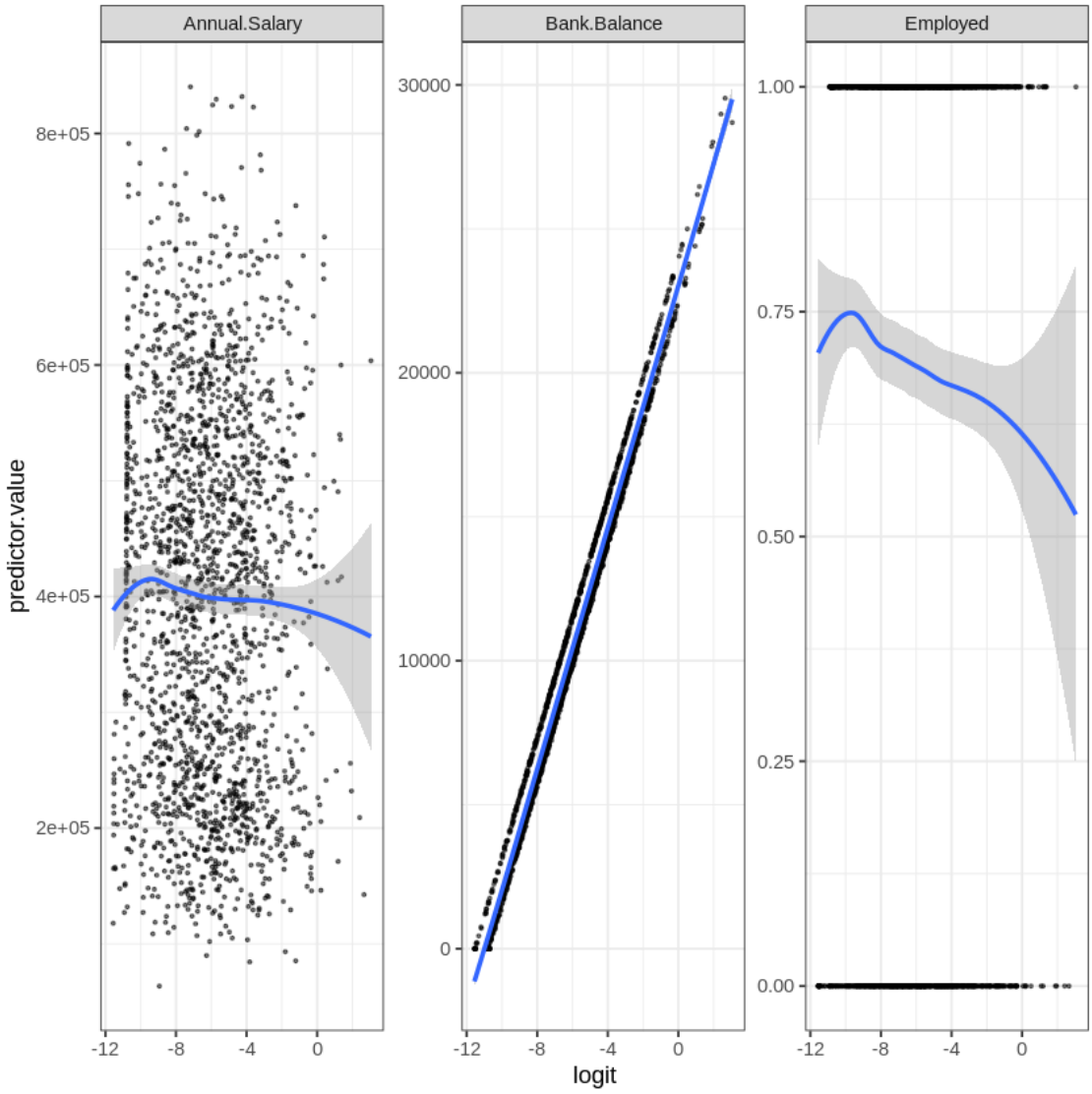
Aqui, o gráfico *Residuals vs Fitted* apresenta uma relação linear entre as variáveis descritivas e a variável objetiva. Além disso, podemos notar no gráfico *Normal Q-Q* que nossos dados não possuem uma distribuição normal. O que não é um problema em nosso caso.

Agora veremos, atributo por atributo, a linearidade da variável com nosso a variável objetivo na escala logit.

[ ]:
```
# tomando dados de predição
probabilities_logit <- model_logit %>% predict(test.data, type = "response")
predicted.classes_logit <- ifelse(probabilities_logit > 0.5, 1, 0)
```

```
[ ]: # tomando subconjunto sem variável alvo
mydata_logit <- test.data[ , !(names(test.data) %in% c('Defaulted.'))]
predictors <-colnames(mydata_logit) # nome das colunas
mydata_logit <- mydata_logit %>% mutate(logit = log(probabilities_logit/
 ↪(1-probabilities_logit))) %>%gather(key ="predictors", value ="predictor.
 ↪value", -logit) # definindo escala logit
ggplot(mydata_logit, aes(logit, predictor.value))+geom_point(size =0.5, alpha␣
 ↪=0.5) +geom_smooth(method ="loess") +theme_bw() +facet_wrap(~predictors,␣
 ↪scales ="free_y") # o plot
```

```
`geom_smooth()` using formula 'y ~ x'
```

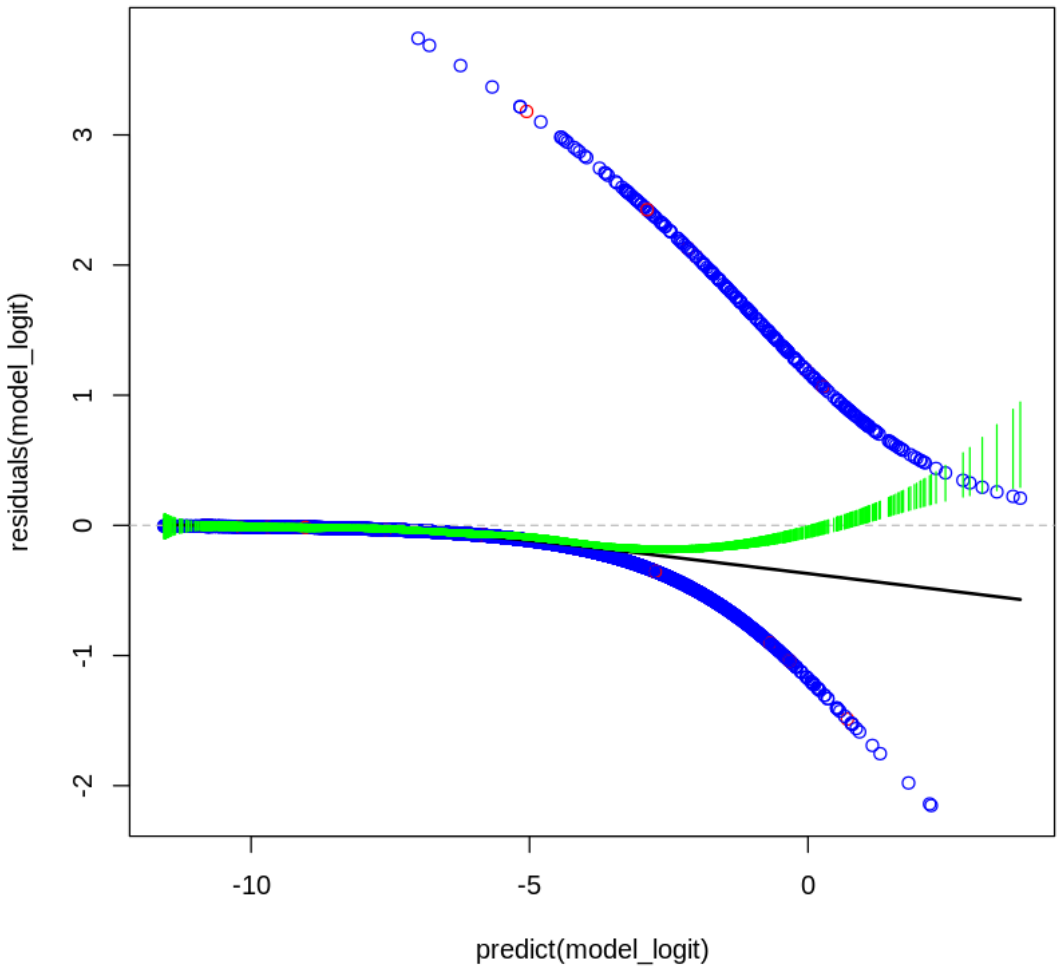


Aqui, pode-se dizer que a variável *Bank Balance* possui uma relação linear com a variável de saída

Defaulted. na escala logit, enquanto as outras duas variáveis não apresentam uma linearidade tão acentuada.

Abaixo, podemos ver a plotagem dos resíduos obtidos no modelo.

[ ]:
```
# plot nos residuos
plot(predict(model_logit),residuals(model_logit),col=c("blue","red")[1+df$Defaulted.
 ↪])
abline(h=0,lty=2,col="grey")
# plot na regressão
lines(lowess(predict(model_logit),residuals(model_logit)),col="black",lwd=2)
# plot do intervalo de confiança
rl=lm(residuals(model_logit)~ splines::bs(predict(model_logit),8))
y=predict(rl,se=TRUE)
segments(predict(model_logit),y$fit+2*y$se.
 ↪fit,predict(model_logit),y$fit-2*y$se.fit,col="green")
```



Em um cenário ideal, teríamos a regressão alinhada com o pontilhado, ao menos que alinhada a partir do intervalo de confiança. De qualquer forma, daremos continuidade nas nossas análises.

Tomaremos agora as métricas de acurácia, de AUC ROC e de F1.

[ ]:
```
# acurácia
acc_logit <- MLmetrics::Accuracy(test.data$Defaulted., predicted.classes_logit)
acc_logit
```

0.974

[ ]:
```
# AUC ROC
auc_logit <- MLmetrics::AUC(test.data$Defaulted., predicted.classes_logit)
auc_logit
```

0.872963915517107

[ ]:
```
# F1 score
f1_logit <- MLmetrics::F1_Score(test.data$Defaulted., predicted.classes_logit)
f1_logit
```

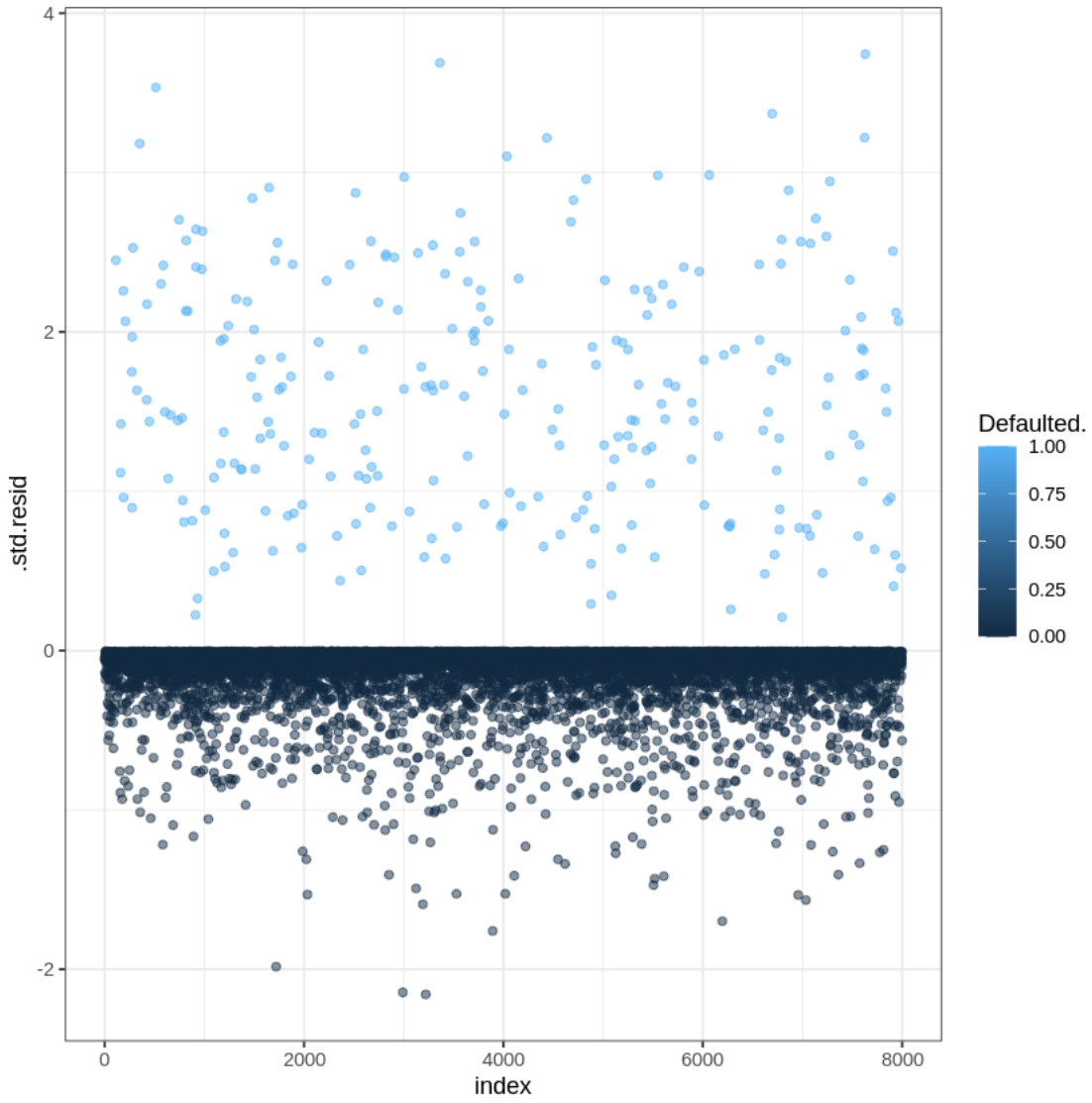
0.986693961105425

Portanto, nosso modelo tem Acurácia de 0.97, AUC de 0.87 e F1_score de 0.98. Valores excelentes para um modelo de classificação.

Vejamos agora os indivíduos influentes na nossa regressão. Sabemos que nem todas as amostras outliers são amostras influentes, portanto podemos realizar a análisa da influência dos dados através da análise de resíduos absolutos padronizados. Quando temos uma amostra com o resíduo padronizado com módulo acima de 3, então este é um ponto influente no conjunto de dados.

[ ]:
```
# tomando os resíduos
std.resid <-augment(model_logit) %>% mutate(index =1:n())
```

[ ]:
```
# plot
ggplot(model.data, aes(index, .std.resid)) +geom_point(aes(color = Defaulted.),␣
↪alpha =.5) +theme_bw()
```



É notavel a presença de algumas amostras com este resíduo padronizado com valor acima de 3, vejamos quais são.

[ ]:
```
out = model.data %>%filter(abs(.std.resid) >3)
out
```

A tibble: 8 × 12

.rownames <chr>	Defaulted. <int>	Employed <int>	Bank.Balance <dbl>	Annual.Salary <dbl>	.fitted <dbl>	.resid <dbl>	.s <
440	1	0	13424.40	262181.3	-5.052163	3.180735	3.
652	1	1	9362.04	619882.4	-6.239255	3.533045	3.
4160	1	1	8382.84	381054.2	-6.798800	3.687795	3.
5015	1	1	12316.32	674185.6	-4.798074	3.100415	3.
5507	1	0	13229.16	208701.4	-5.165902	3.216083	3.
8365	1	0	12158.64	235815.1	-5.670786	3.368746	3.
9523	1	1	11509.92	713222.8	-5.171492	3.217811	3.
9539	1	1	7828.80	553860.5	-7.001240	3.742232	3.

[ ]:
```
out$.rownames
```

1. '440' 2. '652' 3. '4160' 4. '5015' 5. '5507' 6. '8365' 7. '9523' 8. '9539'

Destas amostras que são influentes, todas são inadimplentes.

Vejamos agora qual o desempenho de um modelo de regressão logística, mas desta vez sem o salário anual e sem os pontos amostrais influentes no mmodelo.

3.2 Regressão Logística corrigida

[ ]:
```
test.data <- test.data[-c(as.integer(out$.rownames)),]
train.data <- train.data[-c(as.integer(out$.rownames)),]
```

Abaixo iremos definir o modelo com todas as variáveis descritivas disponíveis no conjunto de dados, exceto o salário anual.

[ ]:
```
model_logit_c <- glm( Defaulted. ~ Employed + Bank.Balance,
                 data = train.data, family = binomial)
```

Uma vez definido o modelo, podemos avaliar como ele se ajustou aos dados. Vejamos agora um resumo dos ajustes.

[ ]:
```
summary(model_logit_c)
```

```
Call:
glm(formula = Defaulted. ~ Employed + Bank.Balance, family = binomial,
    data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1722  -0.1406  -0.0559  -0.0199   3.7498

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.153e+01  4.666e-01 -24.712  < 2e-16 ***
Employed      7.330e-01  1.658e-01   4.422 9.78e-06 ***
Bank.Balance  4.812e-04  2.170e-05  22.174  < 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: 2340.2  on 7994  degrees of freedom
Residual deviance: 1252.5  on 7992  degrees of freedom
AIC: 1258.5

Number of Fisher Scoring iterations: 8
```

Aqui podemos ver que todos os parâmetros são significativos e que este é um modelo melhor ajustado do que aquele apresentado anteriormente, segundo a métrica AIC.

De forma sucinta, veremos o impacto que a remoção das variáveis influentes, e também a remoção da variável descritiva *Annual.Salary* teve no nosso modelo.

[ ]:
```
# tomando dados de predição
probabilities_logit_c <- model_logit_c %>% predict(test.data, type = "response")
predicted.classes_logit_c <- ifelse(probabilities_logit_c > 0.5, 1, 0)
```

[ ]:
```
# acurácia
acc_logit_c <- MLmetrics::Accuracy(test.data$Defaulted., predicted.
↪classes_logit_c)
acc_logit_c
```

0.973973973973974

[ ]:
```
# AUC ROC
auc_logit_c <- MLmetrics::AUC(test.data$Defaulted., predicted.classes)
auc_logit_c
```

0.792095402367216

[ ]:
```
# F1 score
f1_logit_c <-MLmetrics::F1_Score(test.data$Defaulted., predicted.classes)
f1_logit_c
```

0.985335734499614

Portanto, nosso modelo tem Acurácia de 0.97, AUC de 0.87 e F1_score de 0.98. Portanto, houveram mudanças em níveis desconsideráveis no desempenho do modelo. Entretanto, quanto menos dados forem necessários para nosso modelo predizer bem, mais econômico é para os computadores.

3.3 Classificador Random Forest

Como temos uma variável alvo binária, então devemos utilizar Radom Forest como um classificador. Caso contrário, estaríamos usando de forma errônea nossos modelos de aprendizado de máquina.

Inicialmente, precisamos fazer que a variável objetivo seja do tipo Factor.

[ ]:
```
train.data$Defaulted. <- as.factor(train.data$Defaulted.)
test.data$Defaulted. <- as.factor(test.data$Defaulted.)
```

Agora, podemos então processar o modelo. Veja que aqui iremos realizar um processo de validação cruzada no nosso conjunto de dados de treinamento, para assim podermos definir o melhor modelo.

[ ]:
```
set.seed(123)

model_rf <-train(Defaulted. ~ ., data = train.data, method = "rf",trControl␣
↪=trainControl("cv", number = 5),importance =TRUE)
```

```
note: only 2 unique complexity parameters in default grid. Truncating the grid
to 2 .
```

Tenhamos agora uma visão geral do modelo final criado.

[ ]:
```
model_rf$finalModel
```

```
Call:
 randomForest(x = x, y = y, mtry = min(param$mtry, ncol(x)), importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 3.18%
Confusion matrix:
     0  1 class.error
0 7660 68 0.008799172
1  186 81 0.696629213
```

Podemos ver que o modelo com 500 árvores e, logo abaixo, podemos ver a matrix de confusão do modelo (a matriz de confusão para os dados de treinamento).

Agora, podemos seguir para as métricas que já estamos acostumados.

[ ]:
```
predicted.classes <- model_rf %>%predict(test.data)
# acurácia
acc_rf <- MLmetrics::Accuracy(test.data$Defaulted., predicted.classes)
acc_rf
```

0.971471471471472

[ ]:
```
# Auc
auc_rf <- MLmetrics::AUC(test.data$Defaulted., predicted.classes)
auc_rf
```

0.792095402367216

[ ]:
```
# f1
f1_rf <- MLmetrics::F1_Score(test.data$Defaulted., predicted.classes)
f1_rf
```

0.985335734499614

Portanto, o modelo de RF obteve acurácia de 0.97, AUC de 0.79 e F1-score de 0.98.

Abaixo, podemos ver a importância definida para cada variável no modelo.

[ ]:
```
importance(model_rf$finalModel)
```

A matrix: 3 × 4 of type dbl

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
Employed	10.79013	-11.92178	10.250418	5.870551
Bank.Balance	79.45167	134.66013	120.054538	344.037043
Annual.Salary	11.04873	-20.26007	9.354363	164.946438

Em MeanDecreaseAccuracy, podemos ver qual o valor de perda de acurácia caso cada um desses atributos seja removido. Da mesma forma podemos ver para o critério Gini. Uma maneira mais compreensiva de visualizar a importância desses dados é apresentada abaixo.

[ ]:
```
varImp(model_rf)
```

```
rf variable importance

              Importance
Bank.Balance     100.000
Employed           3.618
Annual.Salary      0.000
```

Aqui, temos um valor entre 0 e 100 atribuído para cada variável. Desta maneira, é evidente que o saldo no banco, *Bank.Balance*, é a variável mais relevante, enquanto o salário anual, *Annual.Salary*, é desprezível pro modelo.

3.4 Gráficos adicionais

[ ]:
```
metrics <- data.frame( 'Accuracy'=c(acc_logit, acc_logit_c, acc_rf),␣
↪'AUC'=c(auc_logit, auc_logit_c, auc_rf), 'F1-score'=c(f1_logit, f1_logit_c,␣
↪f1_rf))
rownames(metrics) <- c('logit', 'logit_c', 'rf')
metrics
```

A data.frame: 3 × 3

	Accuracy <dbl>	AUC <dbl>	F1.score <dbl>
logit	0.9740000	0.8729639	0.9866940
logit_c	0.9739740	0.7920954	0.9853357
rf	0.9714715	0.7920954	0.9853357

4 Conclusões e comentários

Agora que já realizamos todas as análises e extraímos as informações necessárias, podemos então discorrer sobre o que foi aqui apresentado. Neste conjunto de dados, temos apenas 3.33% das variáveis classificadas como inadimplentes. Isso somam 333 amostras das 10 mil coletadas. De maneira contra intuitiva o salário anual de um indivíduo não se mostrou como atributo pertinente para a classificação de uma pessoa possivelmente inadimplente. A variável salarial apresentou uma média de aproximadamente 400K de unidades monetárias para pessoas adimplentes e inadimplentes. A baixa correlação, com valor atribuído de -0.02, também é um indicador para podermos pressupor isto antes mesmo de aplicarmos os modelos. A variável que mais se correlaciona com o salário anual é o indicador de emprego , i.e. *Employed*, com indice de 0.75 pontos. O atributo *Employed*, por sua vez, também não apresentou correlação com a inadimplência. Entretanto, a variável de saldo bancário, *Bank Balance*, se mostrou muito eficiente enquanto variável descritiva para a variável alvo de inadimplência. A média do saldo bancário em indivíduo inadimplentes passa de 20.000, enquanto a média do saldo bancário de pessoas adimplentes é ligeiramente abaixo de 10.000. Do conjunto de pessoas adimplentes, existem alguns outliers que passam de 25 mil em saldo bancário, valor que estariam no "bigode"(whisker) do boxplot de inadimplentes.

Ajustando o primeiro modelo de regressão logística, com todas as variáveis e também com todas as amostras, pudemos notar - mais uma vez - que a variável *Annual Salary* não é significante no nosso modelo, isso por ela possuir um P valor de $0.62 > 0.05$, que é o nível significância. Neste modelo, também foi considerado pouco significante a variável *Employed*. No geral, pudemos ver que este primeiro modelo teve acurácia de 0.97, um valor alto o suficiente para aplicações práticas. Para um segundo modelo, removemos não só a variável salarial, como também as amostras influentes no modelo, i.e., aquelas com valor absoluto residual padronizado maior que 3 Neste modelo, todas as variáveis são significativas, entretanto não houve otimização no desempenho, onde obtivemos novamente 0.97 de acurácia. Por fim, o modelo de classificação Random Forest obteve 0.97 de acurácia, além disso, também atribui-se que a variável de salário anual também não é relevante no processo de classificação.

Como todos os modelos obtiveram resultados muito parecidos em acurácia, observamos que em comparação aos outros dois modelos, o modelo de regressão logistica (com todos atributos e amostras) teve destaque, pois apresentou melhor métrica de auc (0.87).

Dessa forma, podemos dizer que o saldo bancário atual de um indivíduo pode vir a ser um bom indicador de potenciais clientes inadimplentes. Entretanto, as variáveis de condição de emprego e de salário anual não parecem ser interessantes para este tipo de predição. Outras possíveis variáveis positivas para o desenvolvimento de modelos preditivos para este modelo, são aquelas omitidas aqui neste conjunto de dados, ou então variáveis temporais, de forma a que possamos acompanhar o comportamento econômico dos clientes. Com estas diferentes variáveis, poderíamos entender o contexto em que cada cliente está inserido sócio-economicamente, e podemos a partir disto buscar por modelos mais sofisticados na área.

5 Referências bibliográficas

[1] Camargos, Marcos Antônio de, Mirela Castro Santos Camargos, Flávio Wagner Silva, Fabiana Soares dos Santos, and Paulo Junio Rodrigues. "Fatores condicionantes de inadimplência em processos de concessão de crédito a micro e pequenas empresas do Estado de Minas Gerais." Revista de Administração Contemporânea 14, no. 2 (2010): 333-352.

[2] Daros, Mariane, and Nelson Guilherme Machado Pinto. “Inadimplência no brasil: uma análise das evidências empíricas.” Revista de Administração IMED 7, no. 1 (2017): 208-229.

[3] Annibal, Clodoaldo Aparecido. Inadimplência do Setor Bancário Brasileiro: uma avaliação de suas medidas. No. 192. 2009.