Andreson Almeida Azevedo

01 de agosto de 2022

Q.1.1: Qual o percentual de faturas emitidas por mês no qual os clientes não pagaram a fatura anterior ?

```
dados <- data.table::fread('data-raw/Questão 1 - Base.txt',sep2 = " ")

'%>%' <- magrittr::'%>%'

dados %>%
  dplyr::mutate(mes_num = lubridate::month(DT_VENCIMENTO),
                MES = toupper(format(DT_VENCIMENTO,"%b"))) %>%
  dplyr::group_by(mes_num, MES) %>%
  dplyr::summarise(TT_FATURAS = length(ID_CONTA),
                   TT_ANT_NPAGAS = sum(ifelse(DS_ROLAGEM == 'FX1',1,0)),
                   '%' = TT_ANT_NPAGAS/TT_FATURAS*100) %>%
  dplyr::arrange(mes_num) %>%
  dplyr::ungroup() %>%
  dplyr::select(c(MES,TT_FATURAS, TT_ANT_NPAGAS,'%')) %>%
  knitr::kable(format = "html", digits = 2, align = "c") %>%
  kableExtra::kable_classic()
```

MES	TT_FATURAS	TT_ANT_NPAGAS	%
JAN	313383	22835	7.29
FEV	313849	27928	8.90
MAR	307318	33432	10.88
ABR	302330	25380	8.39
MAI	301865	30321	10.04
JUN	304366	25977	8.53
JUL	310894	29889	9.61
AGO	317067	28736	9.06
SET	257177	21955	8.54

Q.1.2: Tendo como referência todos os clientes que tiveram fatura emitida no mês de setembro, gere uma base para esses clientes com os seguintes valores calculados:

- **Total de faturas emitidas** nos últimos 6 meses (sem contar com a fatura de setembro);

- **O valor médio de fatura** nos últimos 6 meses (sem contar com a fatura de setembro);
- **Quantidade de vezes que ele ficou sem pagar a fatura anterior** nos últimos 6 meses (sem contar com a fatura de setembro).

```
## Coletando os ID_CONTA do mês de setembro
id_setembro <- dados %>%
  dplyr::mutate(mes_num = lubridate::month(DT_VENCIMENTO),
                MES = toupper(format(DT_VENCIMENTO,"%b"))) %>%
  dplyr::filter(MES == 'SET') %>%
  dplyr::select(ID_CONTA)

## Calculando as métricas para os clientes com fatura emitida no mês de setembro

novas_variaveis <- dados %>%
  dplyr::filter(ID_CONTA %in% id_setembro$ID_CONTA
  ) %>%
  dplyr::mutate(mes_num = lubridate::month(DT_VENCIMENTO),
                MES = toupper(format(DT_VENCIMENTO,"%b"))) %>%
  dplyr::filter(mes_num >= max(mes_num)-6 & mes_num < 9) %>%
  dplyr::group_by(ID_CONTA) %>%
  dplyr::summarise(QTD_FATURAS_ULT_6M = length(ID_CONTA),
                   VL_MEDIO_FATURA = mean(VL_FATURA),
                   QTD_FATURAS_ULT_6M_FX1 = sum(ifelse(DS_ROLAGEM == 'FX1',1,0))
                   )


## coletando usuários que tiveram fatura emitida em setembro
dados_set <- dados %>%
  dplyr::mutate(mes_num = lubridate::month(DT_VENCIMENTO),
                MES = toupper(format(DT_VENCIMENTO,"%b"))) %>%
  dplyr::filter(MES == 'SET')


dados_final <- dados_set %>%
  dplyr::select(ID_CONTA,DS_ROLAGEM,DT_VENCIMENTO) %>%
  dplyr::inner_join(
    novas_variaveis,
    by = "ID_CONTA"
  )

# salvando base de dados no pasta 'data'
data.table::fwrite(dados_final, file = 'data/base_usuarios_setembro.txt')

# apresentando as primeiras linhas da nova base

dados_final %>%
  head() %>%
  knitr::kable(format = "html", digits = 2, align = "c") %>%
  kableExtra::kable_classic()
```

ID_CONTA	DS_ROLAGEM	DT_VENCIMENTO	QTD_FATURAS_ULT_6M	VL_MEDIO_FATURA	QTD_FATURAS_ULT_6M_FX
99416	FX0	2019-09-01	6	2791.62	0
52706	FX0	2019-09-01	6	309.64	0
1024221	FX0	2019-09-01	6	295.18	0
1024238	FX0	2019-09-01	6	870.56	0

ID_CONTA	DS_ROLAGEM	DT_VENCIMENTO	QTD_FATURAS_ULT_6M	VL_MEDIO_FATURA	QTD_FATURAS_ULT_6M_FX1
1024293	FX0	2019-09-01	6	1646.75	0
1024509	FX0	2019-09-01	6	511.90	1

Q.1.3: Utilizando como referência a base calculada na questão anterior, identifiquei qual das 3 variáveis calculadas tem o maior potêncial de preditivo em relação a variável DS_ROLAGEM do mês de setembro

```
# Calculando valor
dados_final %>%
  dplyr::mutate(RESPOSTA = ifelse(DS_ROLAGEM == "FX0",0,1)) %>%
  dplyr::select(-c(ID_CONTA,DT_VENCIMENTO,DS_ROLAGEM)) %>%
  cor() %>%
  knitr::kable(format = "html", digits = 2, align = "c") %>%
  kableExtra::kable_classic()
```

	QTD_FATURAS_ULT_6M	VL_MEDIO_FATURA	QTD_FATURAS_ULT_6M_FX1	RESPOSTA
QTD_FATURAS_ULT_6M	1.00	0.31	0.15	-0.04
VL_MEDIO_FATURA	0.31	1.00	0.06	-0.01
QTD_FATURAS_ULT_6M_FX1	0.15	0.06	1.00	0.19
RESPOSTA	-0.04	-0.01	0.19	1.00

Por meio do coeficiente de correlação de pearson apresentado na tabela anterior, é possível notar que das 3 varáveis calculadas, a variável **QTD_FATURAS_ULT_6M_FX1** é a que tem uma associação mais forte com a variável **DS_ROLAGEM**, indicio de que esta variável tem maior potêncial preditivo que as demais.

Q.2.1: Qual o percentual de adesão mensal por faixa de atraso (Histórico) ?

```
dados <- data.table::fread('data-raw/Questão 2 - Base 1.txt',sep2 = " ")

dados %>%
  dplyr::mutate(
    faixa_atraso = cut(NU_DIAS_ATRASO, 12)
  ) %>%
  dplyr::mutate(mes_num = lubridate::month(DT_ACORDO),
                MES = paste(format(DT_ACORDO,"%b"),format(DT_ACORDO,"%y"),sep="/")) %>%
  dplyr::group_by(faixa_atraso, mes_num, MES) %>%
  dplyr::summarise(
    n = length(RESPOSTA),
    adesao = sum(RESPOSTA, na.rm = TRUE),
    '%' = adesao/n*100) %>%
    dplyr::arrange(mes_num) %>%
    dplyr::ungroup() %>%
  dplyr::select(-c(n,adesao, mes_num)) %>%
  tidyr::spread(key = "MES", value = "%") %>%
  dplyr::select(faixa_atraso,`nov/18`, `mar/19`, `abr/19`, `jun/19`) %>%
  knitr::kable(format = "html", digits = 2, align = "c") %>%
  kableExtra::kable_classic()
```

faixa_atraso	nov/18	mar/19	abr/19	jun/19

faixa_atraso	nov/18	mar/19	abr/19	jun/19
(181,211]	3.56	2.24	26.27	3.82
(211,241]	3.44	1.97	33.60	4.06
(241,271]	3.44	1.26	31.43	3.83
(271,301]	3.17	1.76	36.56	3.19
(301,331]	3.26	1.19	39.02	3.47
(331,360]	2.38	1.15	38.27	2.49
(360,390]	NA	1.35	37.04	3.07
(390,420]	NA	1.09	27.66	2.36
(420,450]	NA	1.80	43.18	2.28
(450,480]	NA	1.41	37.50	2.31
(480,510]	NA	0.94	54.17	2.12
(510,540]	NA	0.76	42.86	1.50

Na tabela anterior é possível observar que em geral quanto maior a faixa de atraso menor a taxa de acordo. Também é possível observar que as taxas de adesão em abril foram muito superiores aos demais meses como pode ser visto na tabela a seguir. Seria interessante investigar se o que aconteceu em Abril e enteder se pode ser reproduzido em próximas campanhas.

```
dados %>%
  dplyr::mutate(
    faixa_atraso = cut(NU_DIAS_ATRASO, 12)
  ) %>%
  dplyr::mutate(mes_num = lubridate::month(DT_ACORDO),
                MES = paste(format(DT_ACORDO,"%b"),format(DT_ACORDO,"%y"),sep="/")) %>%
  dplyr::group_by(mes_num, MES) %>%
  dplyr::summarise(
    n = length(RESPOSTA),
    adesao = sum(RESPOSTA, na.rm = TRUE),
    '%' = adesao/n*100) %>%
  dplyr::ungroup() %>%
  dplyr::arrange(mes_num) %>%
  dplyr::select(-mes_num) %>%
  knitr::kable(format = "html", digits = 2, align = "c") %>%
  kableExtra::kable_classic()
```

MES	n	adesao	%
mar/19	22638	310	1.37
abr/19	811	283	34.90
jun/19	31085	965	3.10
nov/18	17999	587	3.26

Q.2.2: Qual o modelo preditivo você utilizaria para traçar uma estratégia objetivando o aumento de adesão de acordos (Descreva a técnica utilizada)

Modelo de regressão logistica

- É um modelo probabilistico que permite a probabilidade associada à ocorrência de um determinado evento dado os valores observados de outras varáveis;
- Esta técnica costuma ser aplicada em marketing, detecção de fraudes, em riscos financeiros e seguros;

Exemplos: Identificação de transações fraudulentas, identificar usuários com maior propensão à responder um contato de marketing. Para o case técnico: **Identificar quais clientes em atraso tem maior propensão a aderir a um acordo**

- Entre as suas principais vantagens de uso, podemos destacar:

1. Facilidade de lidar com varíaveis categóricas independentes;
2. Fornece resultados em termos de probabilidade;
3. Facilidade de classificar indivíduos em categorias;
4. Necessita de um número pequenos de suposições;
5. É um modelo de fácil interpretação, determina o efeito que os coeficientes exercem sobre a chance de um evento ocorrer. Se um coeficiente é estimado:

- Positivo, Aumenta a probabilidade;
- Negativo, Diminui a probabilidade.

6. Como o modelo retorna probabilidades, é necessário a definir uma regra de predição, em geral se utiliza 0.5, ou seja:

- Se $P(Y = 1) > 0,5$, então se classifica Y = 1;
- Se $P(Y = 1) < 0,5$, então se classifica Y = 0.

Contudo, para a definição da técnica, deve sempre se considerar o tipo de estudo, problema e contexto.

Q.2.3: Quais indicadores e ferramentas você utilizaria para avaliar a perfomance/aderência desse modelo? (Descreva os indicadores utilizada)

- Para uma boa estimativa da eficiência do modelo em termos de classificação, é recomendado separar a amostra em duas partes:

1. uma parte para estimação do modelo (Conjunto de treinamento), e
2. outra parte para testar a eficiência da classificação (holdolt sample)

Para avaliação do poder de descriminação do modelo, se utiliza a seuinte tabela, chamada de matriz de confusão.

		Valor Observado	
		Y = 1	Y = 0
Valor Estimado	$\hat{Y} = 1$	VP	FP
	$\hat{Y} = 0$	FN	VN

A partir da tabela acima é possível calcular as métricas de avaliação, para reduzir trabalho de escrita, vou coloca-lás já em termos do problema proposto na questão seguinte

- **Sensibilidade:** é a probabildade de classificar um usuário que vai fazer acordo quando ele realmente faz acordo. Ou seja, dos usuários que fizeram acordo quantos foram corretamente classificados pelo modelo ?

$$S = \frac{VP}{VP + FN}$$

- **Especificidade:** é a probabilidade de classificar um usuário que não vai fazer acordo quando ele realmente não fez acordo. Ou seja, dos que não fizeram acordo, quantos foram corretamente classificados?

$$E = \frac{VN}{VN + FP}$$

- **Valor preditivo positivo (Precision):** Dos usuários classificados como provável acordo quantos foram corretamente identificados ?

$$VPP = \frac{VN}{VN + FP}$$

- **Valor Preditivo Negativo:** Dos usuários classificados como não acordo, quantos foram corretamente identificados ?

$$VPN = \frac{VN}{VN + FN}$$

- **Precisão (Acurácia):** é a proporção de usuários que foram corretamente classificados ou como acordo ou como não acordo.

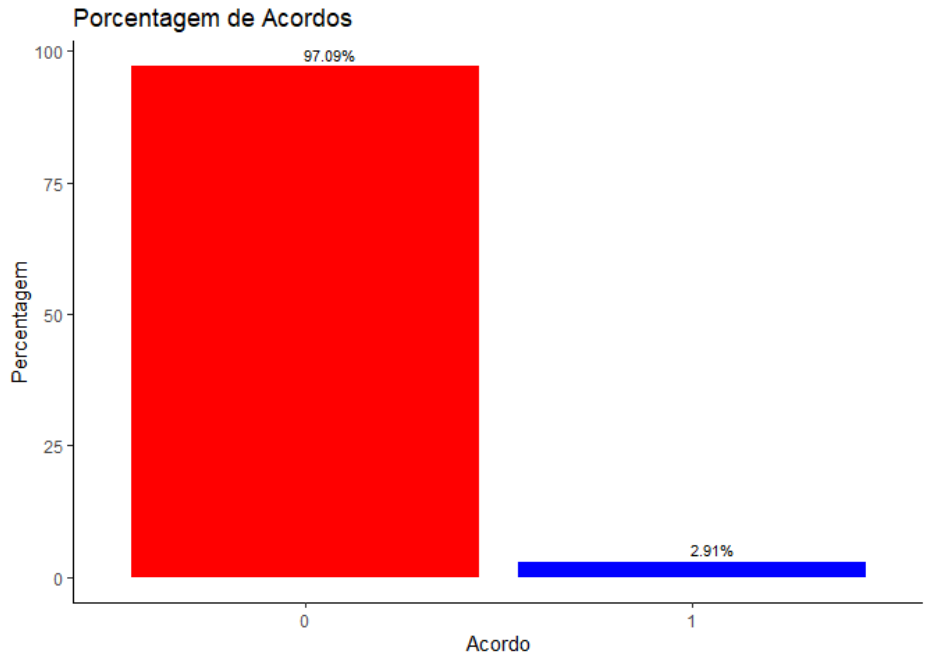
$$ACC = \frac{VP + VN}{VP + VN + FP + FN}$$

Geralmente, sensibilidade alta, significa que estamos classificando muitos usuários como provável acordo. Nessa situação a especificidade é baixa, pois são métricas opostas.

A escolha do modelo então, vai ser feita por meio das métricas acima, em especial sensibilidade e especficidade, em geral a importância dessas métricas depende do problema, em alguns problemas é importante ter um sensibilidade maior, em outros a especificidade. A definição de qual das métricas é mais importante vai depender do custo associado ao erro em cada uma das situações. Para uma definição razoável para o problema é necessário entender com a pessoa de negócio. Para nosso problema, poderia assumir a sensibilidade é mais importante, porque se a sensibilidade for baixa, estou deixando de identificar usuários que iriam aderir o acordo, o que traria uma perda de arrecadação. Mas vale ressaltar que poderia estar correndo risco, visto que não tenho ideia dos custos associados a uma especificidade baixa. Então serei cauteloso, vou buscar a regra de classificação que maximiza a sensibilidade + especificidade.

Q.2.4: Apresente o modelo desenvolvido utilizando a técnica do item (2.2) e as técnicas de avaliação descritas no item (2.3).

Anteriormente, mencionei que para estipular a regra de predição, depende do problema, no nosso caso por exemplo temos uma frequência bem baixa de usuários que fizeram acordo, cerca de 3% dos usuários, um problema de desbalanceamento das classes.



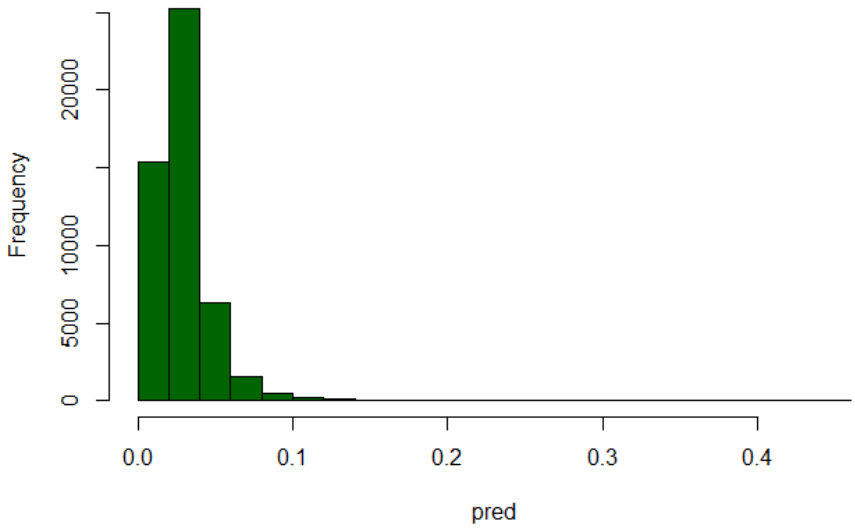
Para o ajuste do modelo, separamos o conjuto de dados em teste (70% dos dados) e treino (30% dos dados), o modelo de regressão logistico foi ajustado e foram selecionadas as variáveis significativas ao nível de significância de 5%, que significa basicamente que é esperado que se coletassemos 20 amostras diferentes da população em 19 delas a estimativa é diferente de zero. Ou seja, que elas contribuem para aumento ou redução da probabilidade de um usuário aceitar o acordo. Os efeitos dos parâmetros são dados em termos das razões de chances (Odds Ratios), para interpretar o efeito percentual de cada variável na probabilidade final só precisa subtrair 1 e multiplicar por 100.

Exemplo: para a variável **NU DIAS ATRASO** $(0.9964 - 1) * 100 = -0.36\%$, ou seja, a cada dia a mais que o usuário permanece em atraso a probabilidade do usuário aderir o acordo se reduz em 0.36%.

	RESPOSTA		
Predictors	*Odds Ratios*	*CI*	*p*
QTD CPC 6M	1.1842	1.1356 – 1.2329	**<0.001**
QTD CP 6M	0.9724	0.9513 – 0.9929	**0.010**
QTD ACIONAMENTO 10D	0.9753	0.9603 – 0.9902	**0.001**
QTD ACIONAMENTO 1M	1.0069	1.0014 – 1.0121	**0.012**
QTD ACIONAMENTO 6M	0.9976	0.9968 – 0.9984	**<0.001**
NU DIAS ATRASO	0.9964	0.9958 – 0.9971	**<0.001**
VALOR CRELIQ	0.9996	0.9995 – 0.9996	**<0.001**
QTD PARCELAMENTO 12M	1.2341	1.1702 – 1.3003	**<0.001**
LIMITE	1.0004	1.0003 – 1.0005	**<0.001**
QTD FX1 6M	1.2110	1.1402 – 1.2856	**<0.001**

Como a frequência de usuários que aceitaram o acordo no conjunto de dados é pequena, as probabilidades calculadas pelo modelo de os usuários aderirem ao acordo são baixas, como pode ser observado no histograma abaixo. Então nessa situação, assumir uma regra de predição, considerando probabilidades maiores que 0.5, vai levar a classificar todos os usuários como não acordo. Uma maneira de evitar isso é

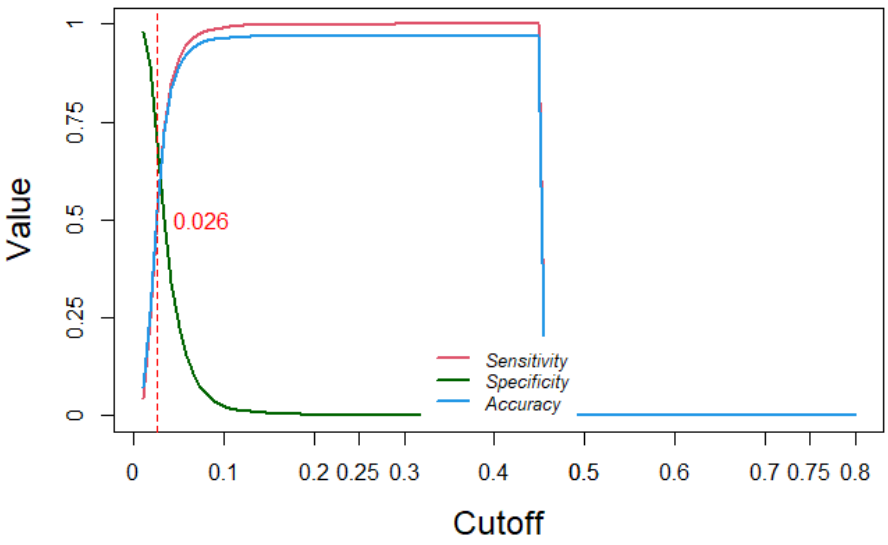
selecionar cortes diferentes de 0.5. Uma alternativa de escolha do ponto de corte, pode ser maximizar alguma daquelas métricas apresentadas anteriormente, para nosso problema iremos maximizar a sensibilidade + especificidade.



Encontrei uma forma para identificação do ponto de corte no *Kaggle* no seguinte notebook feito por Faraz Rahman

```
perform_fn <- function(cutoff)
{
  predicted_churn <- ifelse(pred >= cutoff, "1", "0")
  conf <- confusionMatrix(table(predicted_churn, acordo_train$RESPOSTA))
  accuray <- conf$overall[1]
  sensitivity <- conf$byClass[1]
  specificity <- conf$byClass[2]
  out <- t(as.matrix(c(sensitivity, specificity, accuray)))
  colnames(out) <- c("sensitivity", "specificity", "accuracy")
  return(out)
}


options(repr.plot.width =8, repr.plot.height =6)
#summary(pred)
s = seq(0.01,0.80,length=100)
OUT = matrix(0,100,3)

try(
for(i in 1:100)
{
  OUT[i,] = perform_fn(s[i])
}
, TRUE)

OUT2 <- OUT[1:56,]

cutoff <- s[which(abs(OUT2[,1]-OUT2[,2]) == min(abs(OUT2[,1]-OUT2[,2])))]
#cutoff <- s[which(OUT2[,2] == max(OUT2[,2]))]

plot(s, OUT[,1],xlab="Cutoff",ylab="Value",cex.lab=1.5,cex.axis=1.5,ylim=c(0,1),
```

```
     type="l",lwd=2,axes=FALSE,col=2)
axis(1,seq(0,1,length=5),seq(0,1,length=5),cex.lab=1.5)
axis(2,seq(0,1,length=5),seq(0,1,length=5),cex.lab=1.5)
lines(s,OUT[,2],col="darkgreen",lwd=2)
lines(s,OUT[,3],col=4,lwd=2)
box()
legend("bottom",col=c(2,"darkgreen",4,"darkred"),text.font =3,inset = 0.02,
       box.lty=0,cex = 0.8,
       lwd=c(2,2,2,2),c("Sensitivity","Specificity","Accuracy"))
abline(v = cutoff, col="red", lwd=1, lty=2)
axis(1, at = seq(0.1, 1, by = 0.1))
text(round(cutoff,4), x = cutoff+0.05, y = 0.5, col = "red")
```



O método consiste em interar sobre todas as probabilidades, e obter os valores de sensibilidade e especificidade. Pela abordagem identificamos que o valor do limiar é de 0.026. Definido o limiar, então agora podemos avaliar o modelo ajustado ao conjunto de Treinamento.

```
classe_log <- ifelse(pred > cutoff, "1", "0")

#mean(classe_log != acordo_train$RESPOSTA, na.rm = TRUE) risco

confusionMatrix(table(classe_log, acordo_train$RESPOSTA), positive = "1")
```

```
  Confusion Matrix and Statistics


  classe_log     0     1
           0 24859   433
           1 23044  1009

                 Accuracy : 0.5242
                   95% CI : (0.5198, 0.5286)
      No Information Rate : 0.9708
      P-Value [Acc > NIR] : 1

                    Kappa : 0.0254
```

```
   Mcnemar's Test P-Value : <2e-16

              Sensitivity : 0.69972
              Specificity : 0.51894
           Pos Pred Value : 0.04195
           Neg Pred Value : 0.98288
               Prevalence : 0.02922
           Detection Rate : 0.02045
     Detection Prevalence : 0.48745
        Balanced Accuracy : 0.60933

         'Positive' Class : 1
```

Na tabela acima vemos que o modelo tem uma acurácia de apenas 52.42%, mas a sensitividade de 69.97% é um resultado interessante que traz bons indicios de que o modelo esta classificando bem a classe minoritária. A especificidade ficou em torno de 52%, é interessante entender o impacto de uma especificidade mais baixa, implica entrar em contato com usuários que podem não aderir ao acordo.

Por fim para observar o poder de generalização do modelo, apliquei ao conjunto de testes e novamente utilizei as métricas para avaliá-lo.

O modelo apresentou uma acurácia de apenas 52%, mas lembrando essa métrica mede a precisão que o modelo classificou corretamente os usuários em acordo e não acordo, esse valor mais baixo pode ser explicado pela especificidade mais baixa (52%), ou seja o modelo erra mais em classificar os usuários que não fazem acordo. Em contrapartida, o modelo apresenta uma sensibilidade de quase 72.5%, que significa dizer que a cada 10 usuários que o modelo classificou como acordo, 7 realmente fizeram o acordo. Aqui vale ressaltar que se os custos associados a deixar de apontar usuários que não vão fazer o acordo (usuários podem entrar na próxima campanha para aumento de acordos) forem baixos e/ou o retorno da recuperação dos usuários que fizerem acordos representar um retorno significativo frente aos custos oriundos, faz sentido a utilização do modelo, pois apresenta um bom poder preditivo dos usuários que fizeram acordo.

Se os custos associados ao erro de classificação dos usuários não acordo for alto e/ou o retorno dos usuários não for suficiente para justificar os custos desse erro, é possível que em troca da interpretabilidade (conseguir explicar como cada variável afeta a resposta) podemos melhorar um pouco mais o poder preditivo, fazendo por exemplo a padronização das variáveis associadas a resposta. Outros modelos alternativos também podem ser considerados para este problema: como Análise de discriminante linear e quadrática, suporte vector machines, como também os métodos ensemble (Boosting, Bagging), que são combinações de classificadores para obtenção de um classificar mais poderoso.

```
pred_teste <- predict(logit_final, acordo_test,type = "response")

classe <- ifelse(pred_teste > cutoff, "1", "0")

#mean(classe != acordo_test$RESPOSTA, na.rm=TRUE)

confusionMatrix(table(classe, acordo_test$RESPOSTA), positive = "1")
```

```
  Confusion Matrix and Statistics


  classe     0     1
       0 10697   167
       1  9842   442

                Accuracy : 0.5267
```

```
                 95% CI : (0.52, 0.5335)
    No Information Rate : 0.9712
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0283

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.72578
            Specificity : 0.52081
         Pos Pred Value : 0.04298
         Neg Pred Value : 0.98463
             Prevalence : 0.02880
         Detection Rate : 0.02090
   Detection Prevalence : 0.48629
      Balanced Accuracy : 0.62330

       'Positive' Class : 1
```

Q.2.5: Crie um relatório analítico no Power BI para acompanhar a adesão de acordos e criar insights de como melhorar essa decisão