Reposta para o desafio IEL-CNPq

Raul de Sá Durlo

19/07/2019

No desenvolvimento de modelos de predição, qual a diferença entre as técnicas de regressão linear e regressão logística? Quais são os indicadores para avaliar a performance de aderência do modelo?

Modelos estatísticos

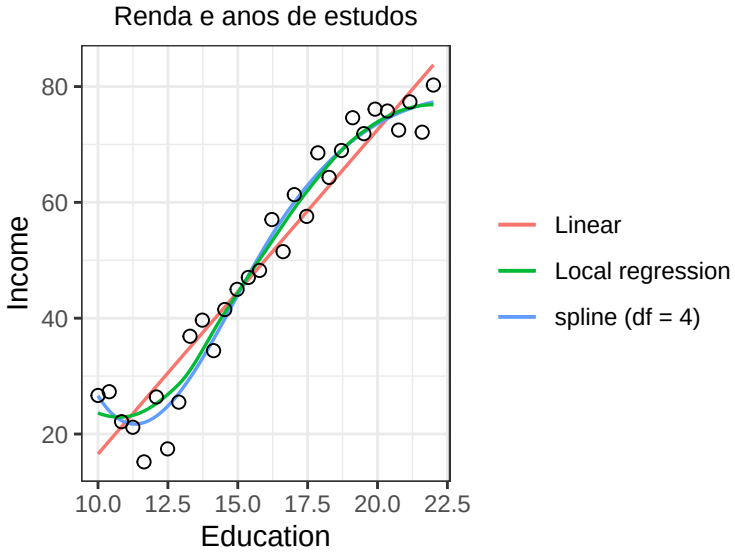
Em um modelo estatístico estamos interessados em obter uma função f que relacione um conjunto de *preditores* X a alguma *variável de resposta* Y. Os preditores $X = (X_1, X_2, ..., X_p)$ também são chamados de *variáveis explicativas*, *variáveis independentes* ou *entrada* (*inputs*). Podemos descrever relação entre X e Y na forma geral:

$$Y = f(X) + \epsilon$$

Onde $f(X)$ representa uma relação sistemática entre o conjunto de preditores X e a variável de resposta Y e ϵ é um termo de erro aleatório independente, com média igual a zero. Desse modo, um modelo pode descrito por:

$$\hat{Y} = \hat{f}(X)$$

O gráfico abaixo mostra uma relação bi-variada entre a renda (Y) de 30 indivíduos com os seus respectivos anos de estudos (X). Cada ponto no gráfico representa um indivíduo (y_1) e as linhas são diferentes modelos estatísticos estimados $(\hat{f})$:



Os modelos representados pelas linhas acima são de diferentes tipos e se ajustam minimizando a distância entre seus valores estimados e os seus valores observados. A forma da curva definida pelo modelo depende de hipóteses assumidas a priori como se há linearidade ou não ou se é paramétrica ou não.

Um modelo paramétrico normalmente apresenta a desvantagem de ser mais simplificador e inflexível, que podem ser atenuados sob a condição de um aumento significativo no número de parâmetros a serem estimados.

Existem também modelos não paramétricos onde não é feita nenhuma suposição sobre a forma funcional de f e, portanto, possui a capacidade de se ajustar melhor ao conjunto de dados.

Por um lado, os modelos mais flexíveis tem maior capacidade de ajuste, porém são mais sensíveis aos erros (*overfitting*). Por outro lado, a simplificação tem a vantagem da interpretabilidade dos dados.

Existem duas razões para se estimar $\hat{f}$, inferencia e predição. A inferência serve para análise da *maneira* de como os preditores $X_1, X_2, ..., X_p$ afetam a variável de resposta Y. Fazemos inferência quando queremos entender a causalidade entre X e Y ou como Y muda em função de $X_1, X_2, ..., X_p$.

O foco na **predição** serve para analisar a precisão de um modelo, ou seja, se os seus valores preditos $\hat{y_0}$ acertariam os valores reais y_0. Mas, como $\hat{f}$ não é, em geral, um estimador perfeito de f, sua diferença (erro) é explicada por fatores *redutíveis* e *irredutíveis*. Podemos decompor esses fatores por meio do quadrado das diferenças entre o valor estimado e a variável de resposta:

$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2 = [f(X) - \hat{f}(X)]^2 + Var(\epsilon)$$

O termo de erro *redutível* corresponde ao termo $[f(X) - \hat{f}(X)]^2$ da equação acima e é aquele decorrente da escolha da forma funcional estimada. Assim, o modelo melhor prediz é aquele que minimiza essa diferença.

Como Y é função de ϵ e, por definição, ϵ não pode ser previsto por X. Algum erro sempre será introduzido ao modelo, daí o termo *irredutível*, denotado por $Var(\epsilon)$.

O Modelo de regressão linear

Um modelo linear assume a função f, linear, em que β_0 e β_1 são os coeficientes a ser estimados.

$$Y \approx \beta_0 + \beta X$$

Os coeficientes do modelo linear podem ser estimados por meio do *método de mínimos quadrados ordinários*

$$min \sum \epsilon^2 = min \sum (Y - \hat{Y})^2 = min \sum [Y - (\beta_0 + \beta X)]$$

$$\hat{\beta} = (X^t X)^{-1} X^t y$$

O gráfico abaixo apresenta a relação entre as vendas (`sales` - em un. de milhar) de um produto em 200 localidades diferentes. O orçamento do departamento de *marketing* é apresentado em unidades de milhar e dividido em três mídias `TV`, `radio` (rádio) e `newspaper` (jornal):

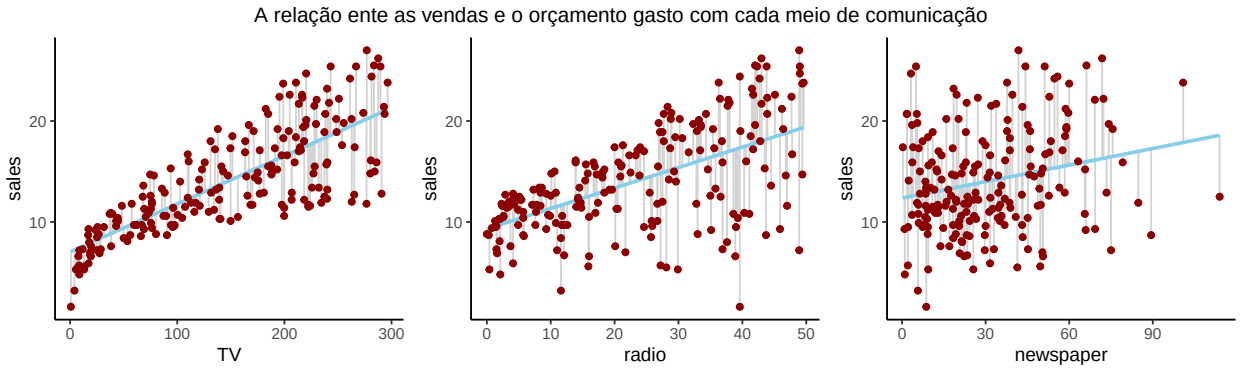


Table 1: Modelo de regressão linear simples: coeficientes de mínimos quadrados ordinários

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.0325935	0.4578429	15.36028	0
TV	0.0475366	0.0026906	17.66763	0

Com base nos gráficos acima podemos observar que os gastos com as mídias parecem corresponder à uma relação linear com as vendas. É de se esperar que, quanto maior os investimentos em propaganda, maiores serão as vendas.

Essa relação linear está representada pela linha reta crescente em azul, que possui inclinação $\hat{\beta}_1$ e intercepto $\hat{\beta}_0$. Essa reta passa por todos os valores preditos pelo modelo e os segmentos em cinza ligam os valores observados aos valores preditos, essa distância representa o termo de erro (ϵ).

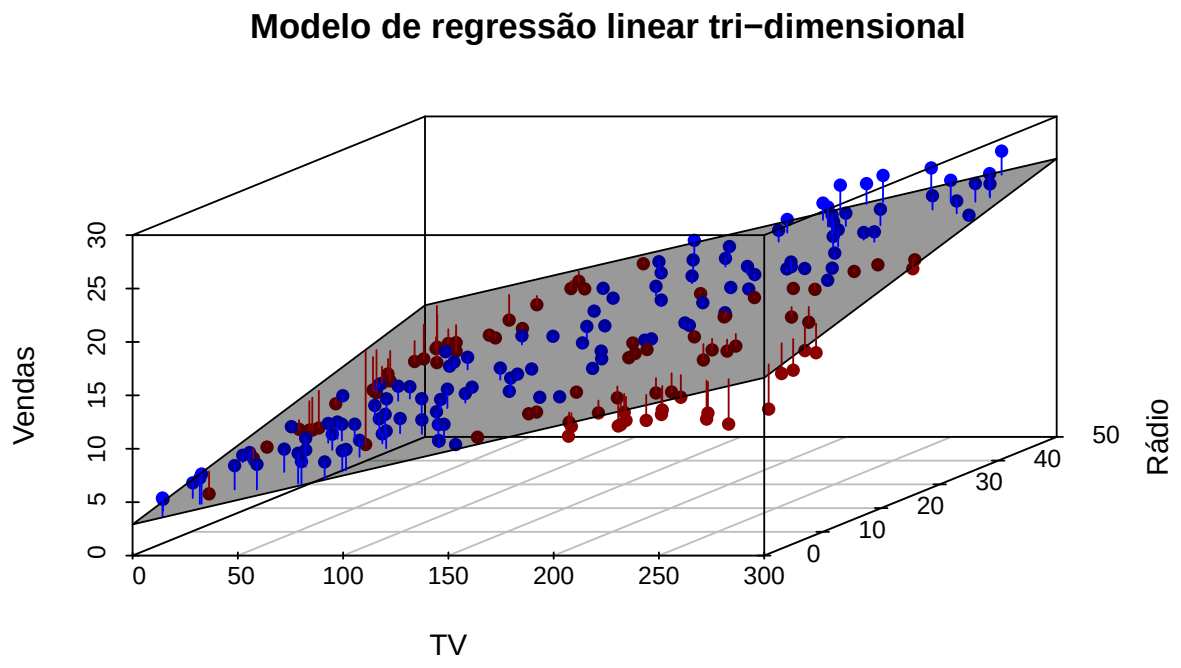
Além da suposição de linearidade do modelo, assume-se também as hipóteses de que $Cov(\epsilon_i, \epsilon_j) = 0$, $\epsilon \sim N(0, \sigma^2)$ e $Y_i \sim N(\beta_0 + \beta X, \sigma^2)$.

$$\texttt{sal\hat{e}s}_i = 7.03 + 0.046 \times \texttt{TV}$$

$$\texttt{sal\hat{e}s}_i = 9.31 + 0.203 \times \texttt{radio}$$

$$\texttt{sal\hat{e}s}_i = 12.35 + 0.055 \times \texttt{newspaper}$$

Outra característica dos modelos apresentados é que eles apresentam somente um parâmetro, sendo considerados, portanto, modelos de regressão linear simples. O modelo linear simples pode ser extendido para o **modelo de regressão linear múltipla**. O gráfico abaixo combina os efeitos de `TV` e `radio` sobre as vendas:



As vendas em função de investimentos em publicidade no rádio e na TV

A tabela abaixo extende o modelo para mais de três variáveis. Na tabela os coeficientes são analisados testando-se a hipótese nula $H_0 : \beta_i = 0$ contra a hipótese alternativa $H_a : \beta_i \neq 0$

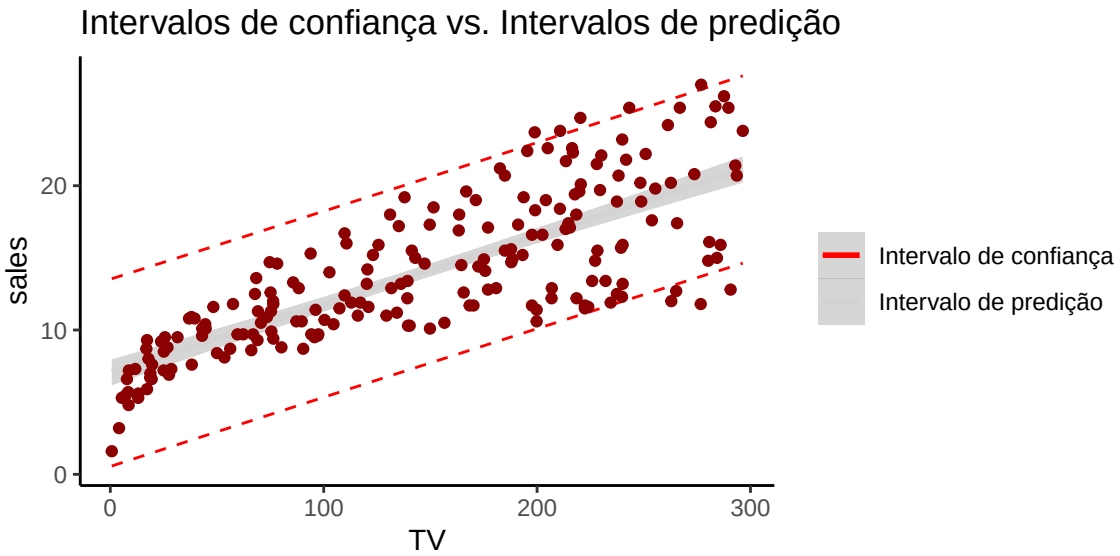
	(1)
(Intercept)	2.939 ***
	(0.312)
TV	0.046 ***
	(0.001)
radio	0.189 ***
	(0.009)
newspaper	-0.001
	(0.006)
N	200
R2	0.897
logLik	-386.181
AIC	782.362

*** p < 0.001; ** p < 0.01; * p < 0.05.

No caso apresentado, a hipótese nula só não é rejeitada para a variável dependente `newspaper`. isso implica que esta variável não é adequada para explicar `sales`. A estatística F testa todos os parâmetros do modelo em conjunto contra a hipótese de modelo nulo (H_a : *pelo menos um* $\beta_i \neq 0$). O R^2 varia entre 0 e 1 e mede o quanto da variação de Y pode ser explicada pelo modelo.

A estratégia para seleção de melhores modelos preditivos parte da estatística F, pois se o modelo é válido então pelo menos 1 dos parâmetros servem para explicar a variável de resposta. Os critérios de seleção são variados e podem partir de um modelo nulo (*backward*) ou de um modelo com todas as variáveis em potencial (*forward*).

Por fim, cabe ressaltar sobre a diferença entre **intervalos de confiança** e **intervalos de predição**: no primeiro caso, os intervalos de confiança testam hipóteses relativas ao modelo em geral, por isso são focados no termo de erro redutível explicados no início deste texto. Já os intervalos de predição são utilizados para predição de um valor pontual.



A forma aditiva ($\beta_0 + \beta_1 X_1 + ... + \beta_p X_p$) é normalmente assumida por questões de simplicidade mas o modelo de regressão linear múltipla pode ser extendido com efeitos de interação ou para preditores qualitativos, como no exemplo abaixo:

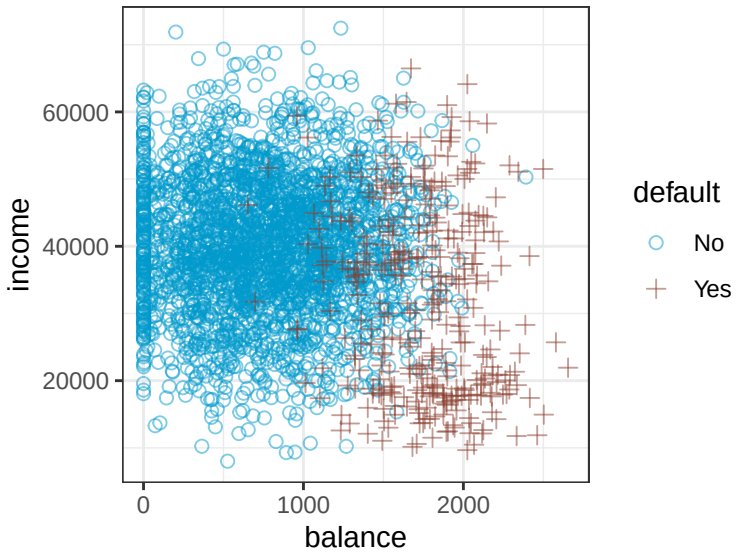
$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income} \begin{cases} \beta_2 \text{ se } \texttt{student} \\ 0 \text{ se } \texttt{nonstudent} \end{cases}$$

Para casos em que a variável qualitativa assume valores qualitativos são formulados os modelos de classificação como por exemplo o *Modelo de Regressão Logística*. A seguir sera apresentado o modelo.

Regressão logística

Apesar de a regressão logística se comportar como uma regressão linear ela não apresenta resposta quantitava. O modelo de regressão logística é adequado para variáveis *qualitativas* ou *categóricas*. A predição de variáveis categóricas é denominada **classificação**.

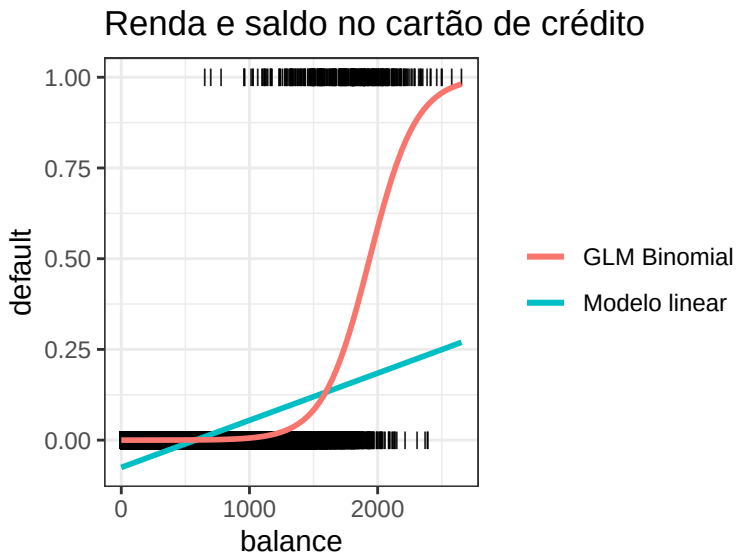
O dado abaixo é denominado `Default`, que contém dados de clientes de um banco onde a variável `balance` é o saldo do cartão de crédito no final de um mês, `income` é a renda média deste cliente e `default` é uma informação binária que indica se o cliente está ou não em débito com o banco.



Aparentemente, clientes devedores (em `default`) são aqueles que costumam apresentar maiores contas no final do mês. Como o fato de o cliente estar ou não em `default`é uma variável binária. Assim, podemos ilustrar nosso problema da seguinte forma:

$$Pf(\texttt{default} = Yes|balance) = p(balance)$$

Na figura abaixo fica claro que o modelo linear não é adequado para este tipo de problema. O modelo prediz probabilidades negativas para `balance` próximos de zero. O modelo linear também é propenso a predizer probabilidades acima de 1 para saldos muito altos. O gráfico abaixo mostra a dispersão entre renda (`income`) e o saldo do cartão de crédito `balance`.



Para contornoar o problema apresentado com o modelo linear, é necessário uma função que retorne resultados entre 0 e 1 para todos os valores de X, como a **função logística**:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

Os parâmetros para estimar a função logística é o de *máxima verossimilhança*. No gráfico acima a linha vermelha evidencia que para valores de saldos muito baixos, a probabilidade de default é muito próxima de zero, mas nunca menor. Manipulando a equação acima, obtemos:

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}$$

O lado esquerdo é denominado *odds* e assume qualquer valor entre 0 e ∞. Usando o logaritmo em ambos os lados, obtemos:

$$log(\frac{p(X)}{1 - p(X)}) = \beta_0 + \beta_1 X$$

Onde $log(\frac{p(X)}{1-p(X)})$ é o *logito* (ou *log-odds*), que por sua vez é linear em X (vide o lado direito da equação anterior. Diferentemente do modelo de regressão linear, os parâmetros são estimados por *máxima verossimilhança* com a *função de verossimilhança*, os coeficientes $\hat{\beta}_0$ e $\hat{\beta}_1$ devem maximizar a equação:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i^t:y_{i^t}=0} (1 - p(x_{i^t}))$$

No modelo abaixo, um incremento em `balance` está recionado a um aumento na probabilidade de `default`, com $\hat{\beta}_1$:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.6513306	0.3611574	-29.49221	0
balance	0.0054989	0.0002204	24.95309	0

Com os coeficientes estimados, é possível predizer a probabilidade de `default` para cada valor de `balance`. Abaixo é demonstrado que para um indivíduo com um saldo de $2000 a probabilidade de ficar inadimplente é de 58.6%.

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.65+0{,}006\times 2000}}{1+e^{-10.65+0{,}006\times 2000}} = 0.586 \; ou \; 58.6\%$$

Performance e aderência

Nesta seção será feita a validação de um modelo de regressão logística multipla. Podemos testar a aderência e a performance dos modelos com a demonstração abaixo realizada no sotware `R`. No primeiro passo, os dados são particionados em uma parte para `treino` e outra parte para `teste`.

```
library(caret)
library(ISLR)
library(tidyverse)
library(lmtest)


Default <- as_tibble(Default)

indice_treino <- createDataPartition(y = Default$default, p = 0.7, list = FALSE)

treino <- Default[indice_treino, ]
teste <- Default[-indice_treino, ]
```

Com os dados de `treino`, utilizaremos a regressão logística para classificação de um modelo com 2 preditores onde a chance de estar em débito com o banco é uma função da renda e do saldo devedor:

```
mod_fit <- train(default ~ balance + income + student,  data=treino, method="glm", family="binomial")
```

Inspecionando os preditores, pode-se perceber que as chances de se estar negativado aumentam em torno de 1 unidade conforme aumentam a renda ou o saldo devedor.

```
exp(coef(mod_fit$finalModel))
##  (Intercept)      balance       income   studentYes
## 2.520825e-05 1.005730e+00 9.999974e-01 4.205393e-01
```

Classificação

```
pred <- predict(mod_fit, newdata = teste)
accuracy <- table(pred, teste$default)
accuracy
##
## pred    No  Yes
##   No  2891   71
##   Yes    9   28
sum(diag(accuracy))/sum(accuracy)
## [1] 0.9733244
```

Matriz de confusão

- **Acurácia:** proporção de predições corretas totais (positivo e negativo)
- **Sensibilidade**: Proporção de verdaeiros positivos
- **Especificidade**: Proporção de verdadeiros negativos
- **Verdadeiro Preditivo Positivo**: Proporção de verdadeiros positivos em relação às predições positivas
- **Verdadeiro Preditivo Negativo**: Proporção de verdadeiros negativos em relação às predições negativas

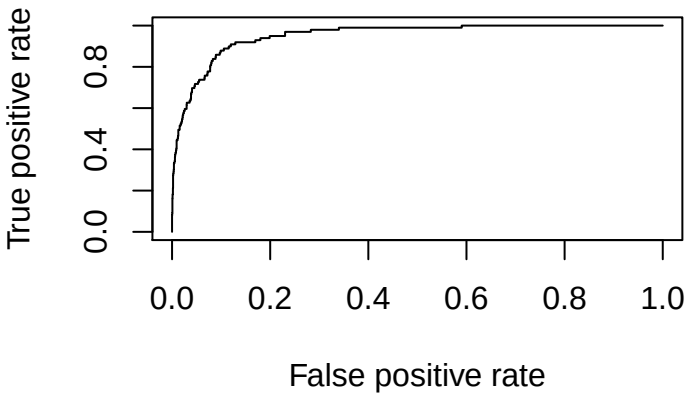
```
pred = predict(mod_fit, newdata=teste)
confusionMatrix(data=pred, teste$default)
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  2891   71
##        Yes    9   28
##
##                Accuracy : 0.9733
##                  95% CI : (0.9669, 0.9788)
##     No Information Rate : 0.967
##     P-Value [Acc > NIR] : 0.02642
##
##                   Kappa : 0.401
##
##  Mcnemar's Test P-Value : 9.104e-12
##
##             Sensitivity : 0.9969
##             Specificity : 0.2828
##          Pos Pred Value : 0.9760
##          Neg Pred Value : 0.7568
##              Prevalence : 0.9670
##          Detection Rate : 0.9640
##    Detection Prevalence : 0.9877
##       Balanced Accuracy : 0.6399
##
##        'Positive' Class : No
##
```

Curva *Receiver Operating Characteristic (ROC)*

A curva ROC plota $P(\hat{Y} = 1|Y = 1)$ (sensibilidade) versus $1 - P(\hat{Y} = 0|Y = 0)$ (1-especificidade) para todos os possíveis pontos de corte entre 0 e 1. Ela mostra o *trade-off* existente entre a qual taxa pode-se predizer corretamente algo contra a taxa de se predizer incorretamente.

```
mod_fit1 <- glm(default ~ balance + income + student,  data=treino, family="binomial")

library(ROCR)
prob <- predict(mod_fit1, newdata=teste, type="response")
pred <- prediction(prob, teste$default)
perf <- performance(pred, measure = "tpr", x.measure = "fpr")
plot(perf)
```




```
auc <- performance(pred, measure = "auc")
auc <- auc@y.values[[1]]
auc
## [1] 0.9517555
```

Conclusão

Nete desafio concluiu-se que modelos de predição devem ser performados atentando-se aos componentes redutíveis e irredutíveis de um modelo de predição. Os modelos de regressão linar são diferentes dos modelos logísticos principalmente em função das suas variáveis de respostas. Os indicadores de aderência do modelo de predição foram apresentados para o MRL (intervalos de confiança e intervalos de predição) e para o modelo logístico com auxílo do software R.

Referência Bibliográfica

- G. James, D. Witten, T. Hastie and R. Tibshirani. **"An Introduction to Statistical Learning, with applications in R"** (Springer, 2013)
- G. James, D. Witten, T. Hastie and R. Tibshirani. ISLR: Data for an Introduction to Statistical Learning with Applications in R. R Package.