Machine Learning

Métodos de Classificação Binária Aplicados à Duas Bases de Dados

Lineu Alberto C. de Freitas
Leonardo Henrique B. Kruger

Fonte dos Problemas

1. IPARDES
2. Classificação do Sexo do Indivíduo pela Voz (Kaggle)

Pacote

Pacote usado - **caret**: **C**lassification **a**nd **Re**gression **T**raining

- **70%** foram separadas para treino
- **30%** foram separadas para teste

Nos dados de treino, na fase de ajuste, fez-se um **5-fold** que foi repetido **5 vezes**

O critério de seleção do melhor modelo foi baseado na área sob a curva ROC (**AUC**)

Métodos

	Modelo	Método caret
1	CART	*rpart*
2	Bagging	*treebag*
3	Random Forest	*ranger*
4	Random Forest Rand.	*extraTrees*
5	Boosting	*adaboost*
6	SVM polynomial kernel	*svmPoly*
7	SVM radial kernel	*svmRadial*
8	SVM linear kernel	*svmLinear*
9	k-Nearest Neighbors	*kknn*
10	glmnet	*glmnet*
11	glm	*glm*
12	An. disc. linear	*lda*
13	An. disc. quadrático	*qda*

Dados IPARDES

Características

1. Número de Linhas - **2394**
2. Número de Colunas - **48**

Variáveis

1. **cid** - cidade
2. **ano**
3. **aaua** - Abastecimento de Água - Unidades Atendidas
4. **nvt** - Nascidos Vivos
5. **abt** - Agências Bancárias
6. **eec** - Consumo de Energia Elétrica (Mwh)
7. **cmp** - Crianças Menores de 2 anos Pesadas
8. **pibpc** - Produto Interno Bruto per Capita
9. **att** - Número de Acidentes de Trânsito
10. **mert** - Matrículas no Ensino Regular
11. **dd** - Densidade Demográfica (hab/km^2)
12. **at** - Área Territorial (km^2)
13. **dsmc** - Distância da Sede Municipal à Capital (km)
14. **fvt** - Frota de Veículos
15. **meit** - Matrículas na Educação Infantil

Variáveis

16. **mct** - Matrículas na Creche
17. **cavt** - Consumo de Água - Volume Faturado (m3)
18. **mpet** - Matrículas na Pré-Escola
19. **vhd** - Vítimas de Homicídio Doloso
20. **meft** - Matrículas no Ensino Fundamental
21. **lat** - Vítimas de Roubo com Resultado de Morte (Latrocínio)
22. **vlc** - Vítimas de Lesão Corporal com Resultado de Morte
23. **vhct** - Vítimas de Homicídio Culposo no Trânsito
24. **mem** - Matrículas no Ensino Médio
25. **ipdm** - Índice Ipardes de Desempenho Municipal (IPDM)
26. **mep** - Matrículas na Educação Profissional
27. **hiv** - Número de Casos por HIV / AIDS
28. **obit** - Óbitos
29. **bcg** - Cobertura Vacinal - BCG (Tuberculose) (%)
30. **hepa** - Cobertura Vacinal - Hepatite A (%)

Variáveis

31. **hepb** - Cobertura Vacinal - Hepatite B (%)
32. **poli**- Cobertura Vacinal - Poliomielite (%)
33. **fa** - Cobertura Vacinal - Febre Amarela (%)
34. **rota** - Cobertura Vacinal - Rotavírus Humano (%)
35. **meni** - Cobertura Vacinal - Meningocócica Conjugada (%)
36. **pne** - Cobertura Vacinal - Pneumocócica 10V (%)
37. **tri** - Cobertura Vacinal - Tríplice Viral (%)
38. **tet** - Cobertura Vacinal - Tetra Viral (%)
39. **dtp** - Cobertura Vacinal - Tríplice (%)
40. **tpb** - Tetra / Penta Bacteriana (%)
41. **pent** - Cobertura Vacinal - Penta Bacteriana ((%)
42. **papi** - Cobertura Vacinal - Papilomavírus Humano (%)
43. **cvdt** - Cobertura Vacinal - Dupla Adulto e Tríplice Acelular Gestante
44. **cvta** - Cobertura Vacinal - Tríplice Acelular Gestante (%)
45. **aaq** - Aeroportos e Aeródromos

Variáveis

- Muitas das variáveis possuiam elevada quantidade de missing (NA)
- Antes da aplicação dos métodos, as variáveis menos NAs foram selecionadas
- Restaram **24** preditoras

Resposta

- **rm** - Receitas Municipais
- **dmt** - Despesas Municipais
- **saldo** - (rm - dmt)
- **class** - risco (saldo $<$ mediana); ok (saldo $>$ mediana)

Objetivo

- **Classificar**, com base em preditoras, um município em duas classes de acordo com o balanço final da cidade.

Resultados

Métodos

	Modelo	Método caret	
1	CART	*rpart*	ok
2	Bagging	*treebag*	ok
3	Random Forest	*ranger*	ok
4	Random Forest Rand.	*extraTrees*	X
5	Boosting	*adaboost*	ok
6	SVM polynomial kernel	*svmPoly*	X
7	SVM radial kernel	*svmRadial*	X
8	SVM linear kernel	*svmLinear*	ok
9	k-Nearest Neighbors	*kknn*	ok
10	glmnet	*glmnet*	ok
11	glm	*glm*	ok
12	An. disc. linear	*lda*	ok
13	An. disc. quadrático	*qda*	ok

Tabela de classificações corretas e incorretas (valores absolutos)

	Modelo	ok;ok	ok;risco	risco;risco	risco;ok
1	CART	283	214	104	50
2	Bagging	192	134	184	141
3	R Forest	197	138	180	136
5	Boosting	190	144	174	143
8	SVMLinear Kernel	160	102	216	173
9	k-Nearest Neighbors	174	152	166	159
10	glmnet	172	116	202	161
11	glm	160	105	213	173
12	An. disc. linear	157	114	204	176
13	An. disc. quadrático	57	26	292	276

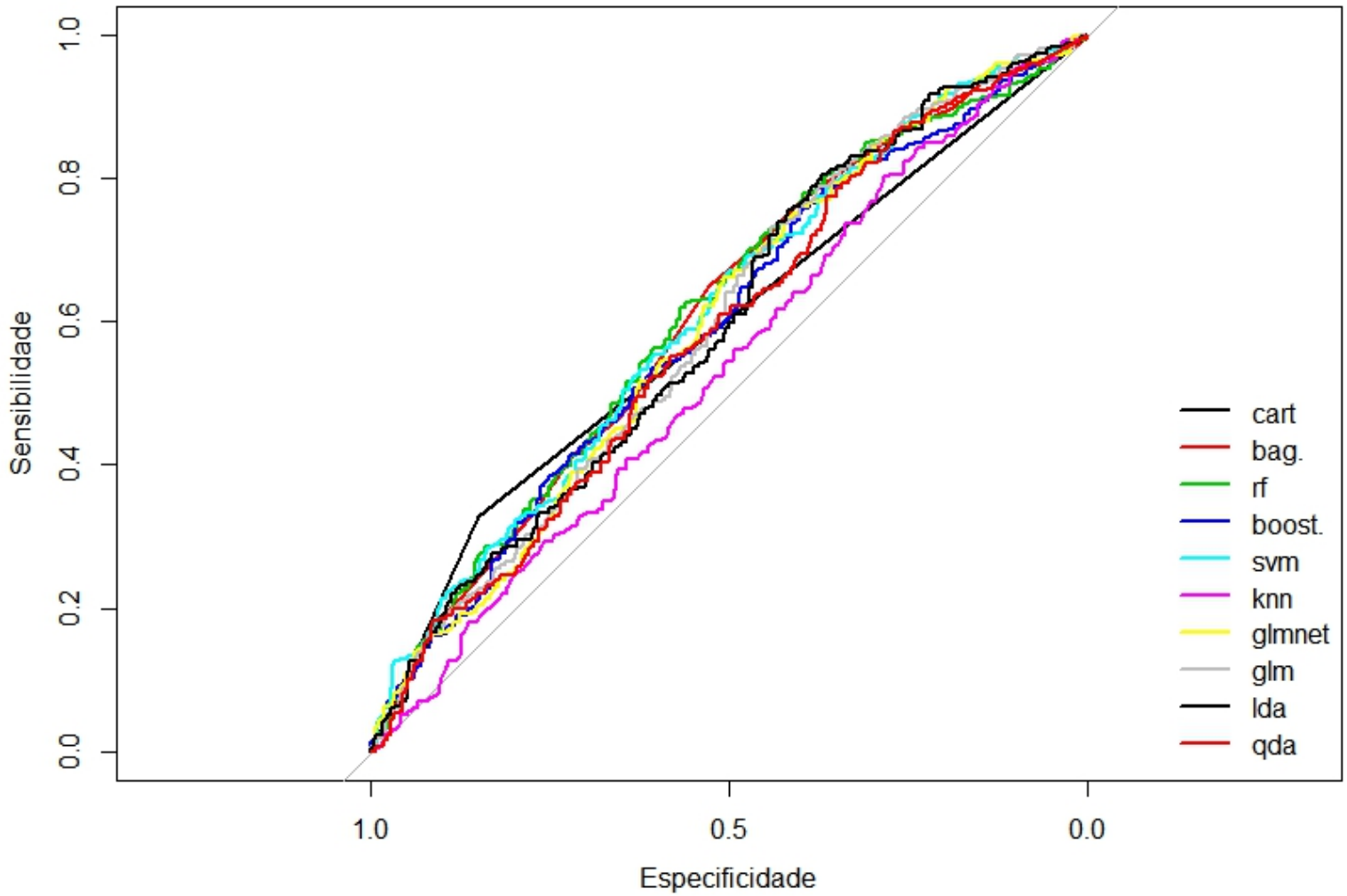
Tabela de classificações corretas e incorretas (proporções)

	Modelo	ok;ok	ok;risco	risco;risco	risco;ok
1	CART	0.435	0.329	0.160	0.077
2	Bagging	0.295	0.206	0.283	0.217
3	R Forest	0.303	0.212	0.276	0.209
5	Boosting	0.292	0.221	0.267	0.220
8	SVMLinear Kernel	0.246	0.157	0.332	0.266
9	k-Nearest Neighbors	0.267	0.233	0.255	0.244
10	glmnet	0.264	0.178	0.310	0.247
11	glm	0.246	0.161	0.327	0.266
12	An. disc. linear	0.241	0.175	0.313	0.270
13	An. disc. quadrático	0.088	0.040	0.449	0.424

Tabela de medidas de qualidade preditiva

	Modelo	Sens.	Espec.	Acur.	AUC
1	CART	0.850	0.327	0.594	0.588
2	Bagging	0.577	0.579	0.578	0.609
3	R Forest	0.592	0.566	0.579	0.610
5	Boosting	0.571	0.547	0.559	0.592
8	SVMLinear Kernel	0.480	0.679	0.578	0.610
9	k-Nearest Neighbors	0.523	0.522	0.522	0.539
10	glmnet	0.517	0.635	0.575	0.597
11	glm	0.480	0.670	0.573	0.596
12	An. disc. linear	0.471	0.642	0.555	0.594
13	An. disc. quadrático	0.171	0.918	0.536	0.582

Curva ROC


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.5
0.0
Especificidade
cart
bag.
rf
boost.
svm
knn
glmnet
glm
lda
qda

Classificação do Sexo do Indivíduo pela Voz (Kaggle)

Características

1. Número de Linhas - **3168**
2. Número de Colunas - **15**

Variáveis

1. **meanfreq** - Frequencia Media (kHz)
2. **sd** - Desvio padrao da Frequencia
3. **sp.ent** - Entropia Espectral
4. **sfm** - Planicidade Espectral (Planeza)
5. **mode** - Frequencia Modal
6. **centroid** - Centroide de Frequencia
7. **peakf** - Frequencia de Pico (Amplitude)
8. **meanfun** - Frequencia Media Fundamental
9. **minfun** - Frequencia Minima Fundamental
10. **maxfun** - Frequencia Maxima Fundamental
11. **meandom** - Frequencia Media Dominante
12. **mindom** - Frequencia Minima Dominante
13. **maxdom** - Frequencia Maxima Dominante
14. **frange** - Amplitude de Frequencia Dominante
15. **modindx** - Indice de Modulacao

Resposta

- **label** - Sexo do indivíduo

Objetivo

- **Classificar**, com base em preditoras, o sexo do indivíduo pelas características vocais.

Resultados

Métodos

	Modelo	Método caret	
1	CART	*rpart*	ok
2	Bagging	*treebag*	ok
3	Random Forest	*ranger*	ok
4	Random Forest Rand.	*extraTrees*	X
5	Boosting	*adaboost*	ok
6	SVM polynomial kernel	*svmPoly*	X
7	SVM radial kernel	*svmRadial*	ok
8	SVM linear kernel	*svmLinear*	ok
9	k-Nearest Neighbors	*kknn*	ok
10	glmnet	*glmnet*	ok
11	glm	*glm*	ok
12	An. disc. linear	*lda*	ok
13	An. disc. quadrático	*qda*	X

Tabela de classificações corretas e incorretas (valores absolutos)

	Modelo	M;M	M;F	F;F	F;M
1	CART	461	28	442	20
2	Bagging	468	18	452	13
3	R Forest	471	13	457	10
5	Boosting	467	16	454	14
7	SVM Radial Basis Kernel	469	19	451	12
8	SVMLinear Kernel	466	13	457	15
9	k-Nearest Neighbors	464	16	454	17
10	glmnet	467	15	455	14
11	glm	467	15	455	14
12	An. disc. linear	454	10	460	27

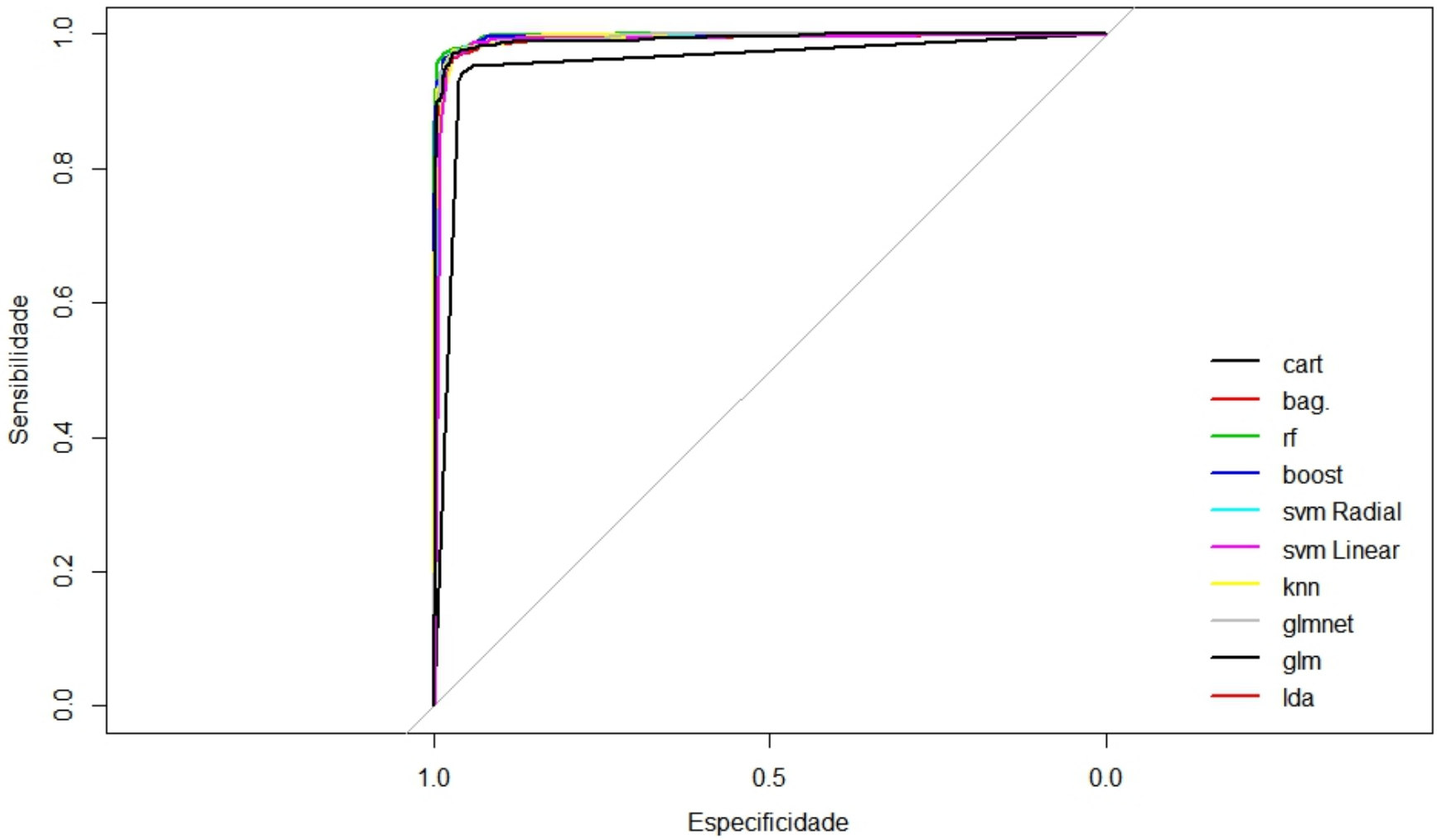
Tabela de classificações corretas e incorretas (proporções)

	Modelo	M;M	M;F	F;F	F;M
1	CART	0.485	0.029	0.465	0.021
2	Bagging	0.492	0.019	0.475	0.014
3	R Forest	0.495	0.014	0.481	0.011
5	Boosting	0.491	0.017	0.477	0.015
7	SVM Radial Basis Kernel	0.493	0.020	0.474	0.013
8	SVMLinear Kernel	0.490	0.014	0.481	0.016
9	k-Nearest Neighbors	0.488	0.017	0.477	0.018
10	glmnet	0.491	0.016	0.478	0.015
11	glm	0.491	0.016	0.478	0.015
12	An. disc. linear	0.477	0.011	0.484	0.028

Tabela de medidas de qualidade preditiva

	Modelo	Sens.	Espec.	Acur.	AUC
1	CART	0.958	0.940	0.950	0.957
2	Bagging	0.973	0.962	0.967	0.990
3	R Forest	0.979	0.972	0.976	0.998
5	Boosting	0.971	0.966	0.968	0.996
7	SVM Radial Basis Kernel	0.975	0.960	0.967	0.994
8	SVMLinear Kernel	0.969	0.972	0.971	0.992
9	k-Nearest Neighbors	0.965	0.966	0.965	0.988
10	glmnet	0.971	0.968	0.970	0.993
11	glm	0.971	0.968	0.970	0.993
12	An. disc. linear	0.944	0.979	0.961	0.991

Curva ROC


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.5
0.0
Especificidade
cart
bag.
rf
boost
svm Radial
svm Linear
knn
glmnet
glm
lda

Considerações finais

Diferenças nos Resultados das Duas Bases

- O que explica um resultado tão ruim em uma base e tão bom em outra (em termos de qualidade preditiva)?

Natureza diferentes das bases

- **IPARDES**: proveniente de consulta online, montada sem conhecimento do problema.
- **Voice**: proveniente de um estudo real.

Natureza das variáveis resposta

- **IPARDES**: variável resposta **contínua** que foi **dicotomizada**
- **Voice**: variável resposta **originalmente com dois níveis**.

Tamanho da base

- **IPARDES**: **2394** linhas
- **Voice**: **3168** linhas

Diferença de 774 observações; diferença que aumentou após os ajustes na base di IPARDES devido aos dados faltantes (NAs)

- **Garbage in, garbage out** (George Fuechsel)

Links de Referência

1. **IPARDES**
2. **Kaggle**
3. **The caret Package**

Dúvidas