oUNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

Lineu Alberto C. De Feitas (GRR20149144)

Leonardo Henrique B. Kruger (GRR20149101)

Métodos de Classificação Binária Aplicados à Duas Bases de Dados

Trabalho referente à disciplina de Machine Learning do Curso de Graduação em Estatística da Universidade Federal do Paraná.

CURITIBA
2017

Sumário

1 INTRODUÇÃO

O objetivo do trabalho foi a aplicação de técnicas de Machine Learning para classificação binária, para isto duas bases de dados foram escolhidas: uma montada via consulta aos dados públicos do IPARDES e outra retirada do site Kaggle.

A primeira base, extraída do site IPARDES, trata de municípios do Paraná dos quais coletou-se uma sequência de variáveis econômicas e sociais. Duas das variáveis coletadas foram utilizadas para a composição de uma variável resposta dicotômica para análise; as variáveis escolhidas foram: a receita municipal e a despesa municipal, a diferença destas gerou a variável saldo que foi dicotomizada em abaixo e acima da mediana e as classes consideradas como município com Risco de Déficit (abaixo da mediana) ou como OK, Sem Risco de Déficit (acima da mediana).

O conjunto de dados retirado do repositório *Kaggle* é um *dataset* onde uma quantidade de áudios foram convertidos em ondas e esses foram analisadas criando-se diversas variáveis sendo uma delas o sexo. O objetivo do estudo foi obter um classificador que diferencia o sexo (masculino ou feminino) com base na composição da onda que cada voz possui.

2 MATERIAL E MÉTODOS

2.1 Material

2.1.1 Conjunto de Dados

O primeiro conjunto de dados utilizado para testar os métodos de Machine Learning é uma base composta por 2394 linhas e 48 variáveis extraídas do IPARDES, porém devido à presença de muitos valores *missing* foi necessário um preprocessamento no sentido de retirar as variáveis com número elevado de dados faltantes afim de obter uma base na qual todos os métodos selecionados fossem aplicáveis; ao final do processo restaram 24 variáveis preditoras.

O segundo conjunto de dados utilizado foi o conjunto de dados com informações sobre as ondas sonoras disponíveis no repositório Kaggle, onde têm-se disponível 15 variáveis e 3168 observações. Para este conjunto, antes da aplicação dos métodos de Machine Learning, realizou-se um preprocessamento no sentido de normalizar as variáveis a serem utilizadas.

2.1.2 Recursos Computacionais

Para ambos os conjuntos de dados foi utilizado o pacote *caret (C*lassification and Regression Training), disponível no software R; as bases foram separadas de forma que 70% dos dados foram utilizados para treino (ajuste) e 30% para teste. Nos dados de treino, fez-se um 5-fold que foi repetido 5 vezes. Adicionalmente, estabeleceu-se para critério de seleção e comparação de modelos a área sob a curva ROC (AUC).

Na Tabela 1 são mostrados os modelos escolhidos e qual o argumento que os descreve nas funções do pacote caret, usado para análise.

Tabela 1: Modelos Utilizados e Nomenclatura Utilizada no Pacote caret

	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*

3 RESULTADOS E DISCUSSÃO

Para o conjunto de dados coletados do IPARDES, alguns modelos não foram aplicados devido a alguma dificuldade técnica. A tabela 2 mostra os métodos aplicados e quais dos métodos foram ou não aplicados.

Tabela 2: Modelos Utilizados, Nomenclatura Utilizada no Pacote caret e Status

	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

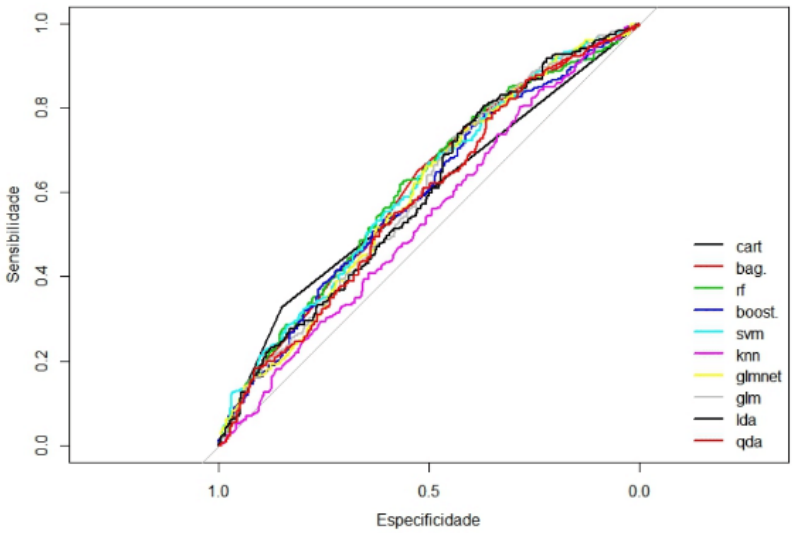
O método Random Forest by Randomization não foi aplicado por dificuldades com as dependências do pacote, os métodos baseados com Support Vector Machine levavam tempo considerável para ajuste do modelo, por isso optou-se pela não aplicação destes métodos.

Tratando da eficiência dos modelos ajustados têm-se os valores obtidos para sensibilidade, especificidade, acurácia e AUC, além do gráfico das curvas ROC dos modelos na Tabela 3 e na Figura 1.

Tabela 3: 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

Figura 1: Curvas ROC



Notou-se que nenhum dos modelos obteve valores para AUC superiores a 0,61. Apontando assim para um desempenho baixo dos modelos na bases utilizadas.

Para o conjunto de dados de origem do *Kaggle,* alguns modelos não foram aplicados devido a alguma dificuldade técnica. A tabela 4 mostra os métodos aplicados e quais dos métodos foram ou não aplicados.

Tabela 4: Modelos Utilizados, Nomenclatura Utilizada no Pacote caret e Status

	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

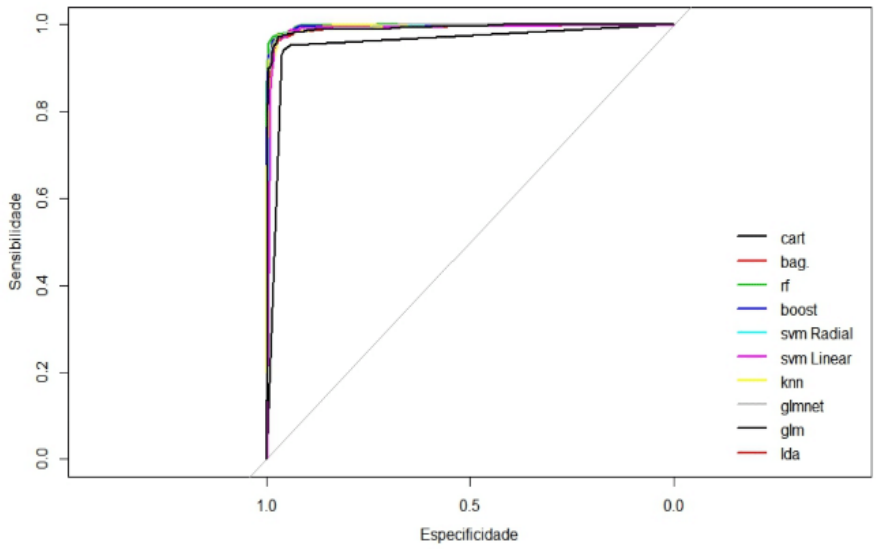
O modelo para análise de discriminante quadrático mostrou performance muito baixa devido problemas no ranqueamento do grupo categorizado como 0, Random Forest apresentou problemas com o formato dos dados e o SVM polynomial excedeu o tempo estabelecido como mínimo para processamento.

Quanto a eficiência dos modelos ajustados temos os valores de sensibilidade, especificidade, acurácia e AUC, além do gráfico das curvas ROC dos modelos na Tabela 5 e na Figura 2:

Tabela 5 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

Figura 2: Curvas ROC



Diferentemente da primeira base de dados analisada, nesta foram alcançados valores satisfatórios de AUC, acurácia, sensibilidade e especificidade, como se pode observar no gráfico da curva ROC que todos os métodos obtiveram bons resultados em termos de qualidade preditiva tendo um máximo de 0.998 no critério de AUC para o modelo de Random Forest.

4 CONCLUSÃO

Uma possível diferença que pode ter influenciado fortemente a diferença entre os resultados da análise das duas bases está possivelmente relacionada à natureza da variável resposta; no primeiro estudo (IPARDES) a variável resposta era de natureza contínua e que foi dicotomizada e no segundo estudo (Kaggle) a variável resposta é originalmente categórica e de dois níveis.

Outro ponto discutido nesse trabalho foi de que a utilização de um número grande de variáveis explicativas não necessariamente vai gerar bons resultados em termos de qualidade preditiva dos modelos, notou-se neste trabalho que a base de dados com número de covariáveis inferior obteve melhor resultado que a base de dados com número maior de covariáveis; o que leva à discussão quanto à qualidade dos dados coletados na base do IPARDES no sentido até que ponto as variáveis selecionadas explicam a variável resposta. Como disse George Fuechsel: *"Garbage in, garbage out"*.

Por fim, ressalta-se também a importância de um o pré-processamento de qualidade nos dados, processo pelo qual são analisadas as variáveis e realizado o tratamento adequado para proporcionar um meio mais propício para o algoritmo ser treinado e obter resultados melhores.