MESTRADO

MATEMÁTICA FINANCEIRA

TRABALHO FINAL DE MESTRADO

DISSERTAÇÃO

PREVISÃO DO INCUMPRIMENTO NO CRÉDITO A EMPRESAS COM CLASSIFICADORES MÚLTIPLOS

ALEXANDRA APARECIDA DELPÓSITO DIAS

SETEMBRO DE 2012

MESTRADO

MATEMÁTICA FINANCEIRA

TRABALHO FINAL DE MESTRADO

DISSERTAÇÃO

PREVISÃO DO INCUMPRIMENTO NO CRÉDITO A EMPRESAS COM CLASSIFICADORES MÚLTIPLOS

ALEXANDRA APARECIDA DELPÓSITO DIAS

ORIENTAÇÃO:

PROF. DOUTOR JOÃO AFONSO RIBEIRO FERREIRA BASTOS

JÚRI:

PRESIDENTE: PROF. DOUTOR ONOFRE ALVES SIMÕES

VOGAL: PROF. DOUTOR JOÃO MANUEL DE SOUSA ANDRADE E SILVA

SETEMBRO DE 2012

Resumo

Neste estudo foram implementados modelos de previsão do incumprimento no crédito a empresas baseados em classificadores múltiplos. O desempenho destes modelos foi comparado com o de classificadores individuais. A capacidade preditiva dos modelos foi avaliada através de curvas ROC e da análise de taxas de erro de classificação. Os resultados sugerem que modelos baseados em classificadores múltiplos têm maior precisão na classificação de incumprimento do que classificadores individuais.

PALAVRAS-CHAVE: Risco de Crédito; Probabilidade de Incumprimento; Regressão Logística; Árvore de decisão; *Bagging*; *Boosting*; *Voting*.

Abstract

This study develops models for predicting credit defaults in the corporate segment using multiple classifiers. The performance of these models was compared with those of individual classifiers. The predictive ability of the competing models was evaluated using ROC curves and error rates of classification. The results suggest that models based on multiple classifiers have a better performance in the classification of credit defaults than individual classifiers.

KEYWORDS: Credit Risk; Probability of Default; Logistic Regression; Decision tree; Bagging; Boosting; Voting.

Agradecimentos

Agradeço ao Professor João Bastos, pela orientação, incentivo e pelos valiosos contributos no desenvolvimento deste trabalho.

Agradeço e dedico todo o meu percurso académico aos meus pais e irmãos, fontes de afecto e segurança.

Ao meu namorado e colega de curso José, que com o seu amor, ajuda, cuidado e compreensão tornou possível o fim desta jornada.

A todos os professores do Instituto Superior Técnico e Instituto Superior de Economia e Gestão, pelos ensinamentos.

À minha amiga, Daniela, que esteve sempre do meu lado, compartilhando e desejando o meu sucesso.

E, principalmente, a Deus, meu grande companheiro, pela força e pelo amparo em todos os momentos da minha vida.

A todos que eu mencionei e a todos que não estão citados, mas que fazem parte do meu mundo relacional, o meu obrigado.

Índice

Lista de Tabelas **viii**

Lista de Figuras **ix**

1 Introdução **1**

2 Revisão de Literatura **4**

3 Descrição e Tratamento dos Dados **6**

3.1 Amostra de Dados 6

3.2 Análise Descritiva 10

3.3 Balanceamento de classes 13

4 Técnicas de previsão do incumprimento **15**

4.1 Regressão Logística 15

4.1.1 Teste de adequação do modelo 16

4.1.2 Significância dos coeficientes do modelo 17

4.1.3 Resultados - Regressão Logística 17

4.2 Árvores de Decisão 20

4.2.1 Algoritmo *REPTree* 20

4.3 Classificadores múltiplos 22

4.3.1 Classificadores múltiplos homogéneos 22

4.3.1.1 *Bagging* 22

4.3.1.2 *Boosting* 23
4.3.2 Classificadores múltiplos heterogéneos 24
4.3.2.1 *Voting* 24

5 Avaliação do poder de previsão **25**
5.1 Curva ROC 25
5.2 Análise dos erros 26

6 Resultados **28**

7 Considerações finais **32**

A Anexos **33**
A.1 Pseudo-Código SMOTE 33
A.2 Pseudo-Código *Bagging* 35
A.3 Pseudo-Código *AdaBoost* 36

Bibliografia **38**

Lista de Tabelas

3.1 Variáveis contabilísticas utilizadas na construção dos rácios económico-financeiros. 7

3.2 Fórmulas de cálculo dos rácios económico-financeiros utilizados na estimação dos modelos. 7

3.3 Análise Descritiva e teste-t para igualdade das médias. 12

4.1 Modelo de previsão do incumprimento dado pela regressão logística. . 18

6.1 Área sob a curva ROC do classificador múltiplo *Voting* para os diferentes métodos de votação. 29

6.2 Área sob a curva ROC dada pelos diferentes tipos de classificador. . . 30

6.3 Taxas de erro dos classificadores dentro da amostra e fora da amostra. 30

6.4 Matriz de classificação dada pelo classificador *Boosting*. 31

Lista de Figuras

4.1 Representação parcial do modelo de previsão do incumprimento obtido pelo algoritmo de árvore de decisão. 21

6.1 Área sob a curva ROC em função do número de árvores de decisão no classificador múltiplo *Bagging*. 28

6.2 Área sob a curva ROC em função do número de árvores de decisão no classificador múltiplo *Boosting*. 29

Capítulo 1

Introdução

Na actual conjuntura económica, a avaliação e controlo do risco de crédito nas instituições bancárias assume especial relevância em virtude da crise sentida no sistema financeiro. O risco de crédito refere-se à possibilidade de um credor incorrer em perdas no empréstimo ou financiamento colocado à disposição de um cliente, mediante o compromisso de pagamento numa data futura. Por forma a ganhar vantagem competitiva em relação à concorrência, os bancos devem implementar um sistema de gestão de risco mais eficaz, tornando indispensável o desenvolvimento de modelos sofisticados de medida do risco de crédito.

O acordo de Basileia II, que consiste num conjunto de regras a serem seguidas pelas instituições bancárias, determina os requisitos mínimos de capital para cobertura dos riscos de crédito, de mercado e operacional. O cálculo do capital regulamentar, k, referente às exposições que não estejam em incumprimento do segmento empresas faz-se através da seguinte expressão:

$$k = LGD \times \left(\Phi \left[\sqrt{\frac{1}{1-\rho}} \Phi^{-1}(PD) + \sqrt{\frac{\rho}{1-\rho}} \Phi^{-1}(0.999) \right] - PD \right) \times \frac{1 + (T - 2.5)b}{1 - 1.5b}$$

onde PD é a probabilidade de incumprimento, LGD é a perda dado o incumprimento, ρ é o coeficiente de correlação entre a rentabilidade da carteira de crédito e o estado geral da economia, $\Phi(.)$ é a função de distribuição normal padrão, T

é uma maturidade, normalmente definida entre 1 e 5 anos, e $b = [0.11852 - 0.05478 \ln(PD)]^2$. De acordo com esta expressão, a probabilidade de incumprimento é um parâmetro fundamental no cálculo dos requisitos mínimos de capital.

Os modelos de classificação normalmente utilizados na previsão do incumprimento consistem em algoritmos como a regressão logística, árvores de decisão, máquinas de vectores de suporte, ou redes neuronais. Estes algoritmos atribuem uma probabilidade de incumprimento, ou pontuação ("score"), a uma empresa ou particular, que indicará o nível de risco da operação. O nível de risco suporta a decisão a tomar pela instituição bancária, sendo que o crédito poderá ser imediatamente aprovado, recusado ou ser sujeito a uma análise mais minuciosa. O nível de risco também determina a taxa de juro associada à operação.

Partindo de um conjunto de observações, estes métodos tentam identificar num espaço de hipóteses a função que melhor soluciona o problema. No entanto, a verdadeira função de representação do problema em análise pode não estar representada no espaço de hipóteses. Pode contornar-se este problema através da estimação de *classificadores múltiplos*. Um classificador múltiplo consiste num conjunto de classificadores, em que as observações são classificadas através da combinação das decisões dos classificadores individuais. Um classificador múltiplo poderá solucionar o problema de representação acima descrito, uma vez que o espaço de hipóteses é expandido. Outro inconveniente dos classificadores individuais é computacional. Muitos destes classificadores utilizam uma forma de pesquisa em que a execução do algoritmo é interrompida ao encontrar um óptimo local. Um classificador múltiplo, construído a partir de diversos pontos no espaço de hipóteses, poderá obter uma melhor aproximação da verdadeira função desconhecida do que qualquer classificador individual.

Neste estudo foram implementados diferentes classificadores múltiplos para previsão do incumprimento no crédito a empresas. Os classificadores múltiplos considerados foram o *Bagging* (Breiman, 1996), o *Boosting* (Freund e Schapire, 1996) e

o *Voting* (Kittler, 1998). O método *Bagging* estima um classificador para cada uma das várias réplicas dos dados obtidas através de amostragem com reposição. A previsão final é obtida por votação maioritária, reduzindo-se desta forma a variabilidade aleatória dos classificadores individuais. O método *Boosting* é um algoritmo iterativo que altera a distribuição das observações de acordo com a classificação anterior. Este algoritmo atribui maior peso às observações classificadas incorrectamente, fazendo com que o algoritmo concentre-se nas observações mais difíceis de classificar. No método *Voting*, a previsão final é obtida através de um esquema de votação de todos os classificadores individuais. Utilizando informação extraída de uma base de dados de uma instituição bancária portuguesa, este estudo mostra que os classificadores múltiplos apresentam melhor capacidade de previsão do incumprimento do que as técnicas tradicionais.

Este estudo desenvolve-se ao longo de 7 capítulos. No capítulo seguinte é realizada uma breve revisão de literatura. O Capítulo 3 descreve a amostra de dados e a análise descritiva das variáveis utilizadas nos modelos. Também é exposto o método de balanceamento de classes considerado neste estudo. Os algoritmos utilizados na previsão do incumprimento são descritos no Capítulo 4. O Capítulo 5 expõe os métodos de avaliação do desempenho dos diferentes algoritmos. No penúltimo capítulo, é comparado o desempenho dos diferentes algoritmos. No último capítulo são apresentadas as considerações finais.

Capítulo 2

Revisão de Literatura

Nas últimas décadas observou-se um enorme desenvolvimento dos modelos de previsão do incumprimento. A ideia de usar rácios financeiros de empresas para prever incumprimentos surge no trabalho seminal de Beaver (1966) e Altman (1968). Beaver (1966) analisou rácios financeiros numa amostra de 79 empresas insolventes, comparando-os com os de outras 79 empresas em situação considerada regular, tendo estas últimas sido seleccionadas através do emparelhamento por indústria e dimensão das empresas insolventes da amostra. A análise univariada dos rácios financeiros mostrou que alguns deles tinham um excelente poder de classificação das empresas insolventes e regulares. A ideia de usar rácios financeiros de empresas num modelo multivariado é proposta em Altman (1968). Neste estudo, foi utilizada uma amostra de 66 empresas (33 solventes e 33 insolventes). Altman usou análise discriminante múltipla para desenvolver um modelo de previsão do incumprimento baseado em cinco rácios financeiros.

Posteriormente, foram realizados vários estudos de previsão do incumprimento baseados na análise discriminante linear. No entanto, a análise discriminante linear foi alvo de críticas devido aos seus pressupostos (Reichert et al., 1983). De facto, esta técnica possui pressupostos bastante restritivos, como a normalidade das variáveis independentes e a igualdade das matrizes de variância-covariância dos gru-

pos de interesse. Estas suposições poderão não ser válidas em muitas situações. A regressão logística é uma técnica de previsão de variáveis binárias que constitui uma alternativa à análise discriminante linear na construção de modelos de previsão do incumprimento. Foram publicados vários estudos de análise do risco de crédito baseados nesta técnica. Por exemplo, Wiginton (1980) comparou o desempenho do método de regressão logística com o método de análise discriminante, concluindo que o modelo de regressão logística tem um desempenho ligeiramente superior.

Mais recentemente, têm sido propostas técnicas não-paramétricas para análise do risco de crédito, entre elas, árvores de decisão, redes neuronais artificiais, modelos de k-vizinhos mais próximos e máquinas de vectores de suporte. O modelo desenvolvido por Frydman et al. (1985) é baseado em árvores de classificação. Este modelo apresentou desempenho superior ao das técnicas baseadas na análise discriminante. Henley e Hand (1996) compararam a regressão logística e as árvores de decisão com o método dos k-vizinhos mais próximos e obtiveram bons resultados com esta técnica não-paramétrica. Baesens et al. (2003) sugerem que as redes neuronais artificiais e as máquinas de vectores de suporte possuem um bom desempenho na classificação de incumprimentos.

A classificação de incumprimentos com redes neuronais artificiais foi abordada em vários estudos (ver, por exemplo, Jensen, 1992; West et al., 2005). Jensen (1992) utilizou as redes neuronais artificiais no desenvolvimento de um modelo de previsão do incumprimento. O modelo considerado foi baseado numa amostra de 125 clientes. Neste trabalho, as redes neuronais artificiais apresentaram um bom desempenho. West et al. (2005) estudaram a aplicação dos classificadores múltiplos *Bagging* e *Boosting* em problemas de análise do risco de crédito. O classificador base considerado foi a rede neuronal artificial. Neste estudo os resultados obtidos pelos classificadores múltiplos superaram os resultados alcançados pela rede neuronal artificial.

Capítulo 3

Descrição e Tratamento dos Dados

3.1 Amostra de Dados

A amostra utilizada neste trabalho consiste na informação extraída de uma base de dados de uma instituição bancária portuguesa. Estes dados são anteriores ao ano de 2008 e reflectem as informações contabilísticas das empresas numa determinada data. Os critérios para análise do crédito são baseados em 18 rácios económico-financeiros obtidos de 4000 empresas, das quais 3886 cumpriram com as suas obrigações contratuais e 114 não cumpriram. Segundo as normas do acordo de Basileia II, é considerado como incumprimento um atraso no pagamento superior a 90 dias.

A escolha de uma metodologia baseada em rácios económico-financeiros prende-se com o facto de ter sido uma das primeiras concebidas para a previsão de incumprimento (Beaver, 1966; Altman, 1968), e a mais utilizada na estimação da probabilidade de incumprimento no segmento empresas. A Tabela 3.1 descreve as variáveis contabilísticas utilizadas na construção dos rácios económico-financeiros. Na Tabela 3.2 podem encontrar-se as fórmulas de cálculo dos rácios económico-financeiros utilizados na estimação dos modelos.

Notação	**Variável contabilística**
A	Amortizações
AAS	Avaliação dos accionistas e sócios
AC	Activo circulante
ATL	Activo total líquido
CDB	Caixas e depósitos bancários
CE	Custos extraordinários
CPT	Capital próprio total
DF	Despesas financeiras
DTCP	Dívidas a terceiros de curto prazo
DTMLP	Dívidas a terceiros de médio e longo prazo
ISRE	Imposto sobre o rendimento do exercício
MLL	Meios libertos líquidos
P	Provisões
PC	Passivo circulante
PE	Proveitos extraordinários
PT	Passivo total
R	Reservas
RC	Resultados correntes
RLE	Resultados líquidos do exercício
RT	Resultados transitados
VPS	Vendas e prestações de serviços

Tabela 3.1: Variáveis contabilísticas utilizadas na construção dos rácios económico-financeiros.

Nome	**Fórmula**
I1	(DTCP+DTMLP-AAS)/VPS
I2	(DTCP-CDB)/ATL
I3	(RC+A)/VPS
I4	(RT+R)/ATL
I5	AC/DTCP
I6	CDB/PC
I7	CPT/ATL
I8	CPT/(DTMLP+DTCP)
I9	DF/VPS
I10	RC/DTCP
I11	RLE/ATL
I12	(RLE+ISRE-PE+CE+A)/DF
I13	(RLE+ISRE-PE+CE+A+P)/PT
I14	RLE/CPT
I15	DTCP/AC
I16	DTCP/PT
I17	DF/PT
I18	MLL/DTCP

Tabela 3.2: Fórmulas de cálculo dos rácios económico-financeiros utilizados na estimação dos modelos.

De seguida, realiza-se uma breve análise dos rácios económico-financeiros (Carvalho e Magalhães, 2002).

I1: Um valor elevado deste rácio é desfavorável para uma empresa, dada a evidência de que as vendas e prestações de serviços não são suficientes para cobrir as dívidas a terceiros de curto, médio e longo prazo.

I2: Quando o activo líquido não cobre o valor das dívidas de curto prazo (descontando os fluxos de caixa e depósitos bancários) existe um aumento na probabilidade de incumprimento. Portanto, quanto maior este rácio, maior é a probabilidade de incumprimento.

I3: O aumento deste rácio contribui para o decréscimo da probabilidade de incumprimento, uma vez que indica um acréscimo na quantia resultante do volume de negócios depois de deduzidos os custos operacionais e adicionada a diferença entre a receita financeira e os encargos financeiros.

I4: Na conta resultados transitados são registados os resultados líquidos e dividendos antecipados, provenientes do exercício anterior. O aumento desta variável, representa um decréscimo na probabilidade de incumprimento.

I5: O rácio de liquidez geral é um rácio financeiro que mede a capacidade da empresa de fazer face às suas responsabilidades de curto prazo. Quanto mais elevado este rácio, maior a solvabilidade de curto prazo da empresa. Quanto mais baixo, maior a vulnerabilidade. Conclui-se assim, que quanto maior este rácio, menor a probabilidade de incumprimento.

I6: O rácio de liquidez imediata é um rácio financeiro que mede a capacidade da empresa de fazer face às suas responsabilidades de curto prazo utilizando apenas a sua disponibilidade financeira imediata. Quanto maior este indicador, menor é a probabilidade de incumprimento.

I7: O rácio de autonomia financeira indica a percentagem do activo que é coberta por capitais próprios. Quanto mais elevado este rácio, maior a estabilidade financeira da empresa.

I8: O rácio de solvabilidade financeira mede a relação entre os capitais próprios e o total do passivo. É, portanto, importante controlar este indicador financeiro por forma a não colocar em causa a continuidade da empresa. Um valor muito baixo deste rácio pode indiciar uma fraca viabilidade da empresa no futuro, pois significa uma elevada fragilidade económico-financeira.

I9: Quanto mais alto for este rácio mais desfavorável será a situação em que a empresa se encontra, dado que significa que as políticas de endividamento da empresa não estão a ser eficazes, sendo necessário que uma maior parte das vendas e prestações de serviços seja utilizada para cobrir os custos financeiros. Logo, este rácio contribui para um acréscimo na probabilidade de incumprimento.

I10: Quanto menores as dívidas ou quanto melhores forem os resultados da empresa maior será a sua estabilidade financeira e consequentemente terá uma diminuição na probabilidade de incumprimento.

I11: A rendibilidade do activo é um indicador económico que mede a capacidade dos activos da empresa em gerar retorno financeiro. Este indicador obtém-se pela divisão dos resultados líquidos pelo valor líquido dos activos da empresa. Um resultado elevado reflecte a elevada capacidade que os activos da empresa têm para gerarem retorno financeiro.

I12: Quanto maiores os resultados operacionais e financeiros, maior será este rácio e isso traduz-se numa menor probabilidade de incumprimento.

I13: Interpretação análoga à do índice anterior.

I14: O rácio de rentabilidade dos capitais próprios é um indicador económico que mede a capacidade dos capitais próprios da empresa em gerar retorno financeiro. Quanto mais elevado for este indicador, mais atraente será a empresa para eventuais investidores e maiores possibilidades a empresa terá de desenvolver a sua actividade futura com recurso ao auto-financiamento.

I15: Os activos circulantes são constituídos por um conjunto de contas do activo que se antecipa serem convertíveis em dinheiro num prazo inferior a um ano. Um

valor alto deste rácio representa acréscimo na probabilidade de incumprimento, já que neste caso existe uma preponderância da variável constante no numerador.

I16: O passivo de uma empresa é constituído pelas seguintes rubricas fundamentais: dívidas a terceiros de médio e longo prazo (abrange todas as dívidas exigíveis num prazo superior a um ano) e dívidas a terceiros de curto prazo (abrange todas as dívidas exigíveis num prazo inferior a um ano). Um valor elevado deste rácio traduz o facto do passivo ser composto maioritariamente por dívidas a terceiros de curto prazo. Logo, quanto maior esta variável maior é a probabilidade de incumprimento.

I17: Quanto maior este rácio maior é a probabilidade de incumprimento, dado que as despesas provenientes da emissão de obrigações e de empréstimos contraídos a curto, médio e longo prazo são elevados.

I18: Quanto mais elevado este rácio, maior a capacidade da empresa para pagamentos de dividendos, reembolso de capital alheio e autofinanciamento.

3.2 Análise Descritiva

A Tabela 3.3 mostra os valores médios dos rácios económico-financeiros para as empresas em situação de incumprimento e em situação regular. Os valores médios dos rácios estão de acordo com a interpretação efectuada na Secção 3.1. Por exemplo, verifica-se que o valor médio do rácio I1 é maior no caso das operações que se encontram em situação de incumprimento (2.4479), do que para as operações regulares (1.5157). Este resultado está de acordo com a interpretação deste rácio e sugere uma maior dificuldade das empresas em situação de incumprimento em efectuarem o pagamento das dívidas, apenas por meio das vendas e prestações de serviços.

Com o objectivo de testar se a diferença entre as médias de cada variável para os clientes em situação de incumprimento e clientes regulares é estatisticamente significativa utilizou-se o teste de t-Student. Este teste requer que as populações possuam distribuição normal e variâncias iguais. Dado que a dimensão da amostra

original é de 4000 clientes, como consequência do teorema do limite central utilizou-se o teste de t-Student no caso de variâncias amostrais semelhantes e o teste de t-Student com correcção de Welch para as situações de variâncias amostrais diferentes. Assim, para as amostras referentes aos créditos em incumprimento (I) e regulares (R) as hipóteses testadas foram:

$$H_0 : \mu_I = \mu_R \quad vs \quad H_1 : \mu_I \neq \mu_R$$

Para testar a homogeneidade das variâncias, foi aplicado o teste de Levene a cada um dos rácios económico-financeiros considerando as seguintes hipóteses:

$$H_0 : \sigma_I^2 = \sigma_R^2 \quad vs \quad H_1 : \sigma_I^2 \neq \sigma_R^2$$

Os desvios-padrão amostrais, os valores das estatísticas de teste e os respectivos valores-p encontram-se na Tabela 3.3. Para uma dimensão de teste de 5%, apenas existem diferenças estatisticamente significativas nos rácios I3, I4, I7, I9, I11 e I13.

Nesta secção, os resultados foram obtidos através do *software* SPSS versão 20. Nos restantes capítulos, utilizou-se a linguagem de programação Java implementada no *software WEKA - Waikato Environment for Knowledge Analysis* versão 3.6.6.

Índice	Estado	Estatísticas		Teste-t	
		Média	Desvio-Padrão	t	valor-p
I1	Crédito Regular	1.5157	2.8878	1.830	0.070
	Crédito em Incumprimento	2.4479	5.4153		
I2	Crédito Regular	0.4828	0.3217	1.740	0.082
	Crédito em Incumprimento	0.5360	0.3372		
I3	Crédito Regular	0.0651	0.1864	-3.997	0.000
	Crédito em Incumprimento	-0.0060	0.2097		
I4	Crédito Regular	0.0464	0.1616	-4.284	0.000
	Crédito em Incumprimento	-0.0196	0.1780		
I5	Crédito Regular	9.9970	231.75	-0.157	0.876
	Crédito em Incumprimento	6.5956	32.386		
I6	Crédito Regular	0.1095	0.1639	-1.502	0.133
	Crédito em Incumprimento	0.0862	0.1458		
I7	Crédito Regular	0.1989	0.2211	-3.481	0.001
	Crédito em Incumprimento	0.1261	0.1829		
I8	Crédito Regular	0.4355	2.2761	-1.042	0.297
	Crédito em Incumprimento	0.2132	0.3285		
I9	Crédito Regular	0.0587	0.1072	2.701	0.008
	Crédito em Incumprimento	0.1353	0.3022		
I10	Crédito Regular	0.4641	21.008	-0.211	0.833
	Crédito em Incumprimento	0.0481	0.6883		
I11	Crédito Regular	0.0162	0.1181	-2.846	0.004
	Crédito em Incumprimento	-0.0156	0.0835		
I12	Crédito Regular	23.222	616.19	-0.375	0.708
	Crédito em Incumprimento	1.5919	9.3826		
I13	Crédito Regular	0.1039	0.2607	-2.867	0.004
	Crédito em Incumprimento	0.0337	0.1226		
I14	Crédito Regular	0.0728	2.3935	-1.183	0.237
	Crédito em Incumprimento	-0.1958	2.2592		
I15	Crédito Regular	0.3800	0.3167	0.830	0.408
	Crédito em Incumprimento	0.4030	0.2909		
I16	Crédito Regular	0.6990	0.3033	0.102	0.919
	Crédito em Incumprimento	0.7020	0.3166		
I17	Crédito Regular	0.0541	0.6962	0.031	0.975
	Crédito em Incumprimento	0.0561	0.0275		
I18	Crédito Regular	0.4835	16.979	-0.220	0.826
	Crédito em Incumprimento	0.1334	0.8979		

Tabela 3.3: Análise Descritiva e teste-t para igualdade das médias.

3.3 Balanceamento de classes

Em muitos problemas de classificação de variáveis com resposta binária, como ocorre quando se desenvolve um modelo de previsão de incumprimento, é observado um desbalanceamento significativo entre as duas classes. Quando o número de elementos entre as classes é desproporcional, as observações da classe minoritária são geralmente classificadas de forma incorrecta, influenciando negativamente o desempenho dos algoritmos de classificação. Uma forma de solucionar este problema consiste em efectuar um balanceamento de classes. Esta técnica pode ser realizada de duas formas: inserir elementos na classe minoritária (*over-sampling*), ou eliminar elementos da classe maioritária (*under-sampling*). A eliminação de elementos da classe maioritária conduz a uma perda de informação que possivelmente resultará num pior desempenho dos modelos estimados.

O desempenho do algoritmo SMOTE pode ser degradado na presença de ruídos, dado que esta técnica fará crescer a região de decisão das observações da classe minoritária e por isso aumentará a capacidade de generalização dos classificadores para estes casos, o que é desejado. No entanto, em amostras com esta característica, poderá gerar, ou mesmo aumentar, a ocorrência de observações indesejáveis.

Neste trabalho, efectuou-se um balanceamento dos dados através do método SMOTE - *Synthetic Minority Over-Sampling* (Chawla et al., 2002). Conforme mostra o pseudo-código no Apêndice 1, este método, em vez de replicar observações da classe minoritária, gera novos casos sintéticos tendo como base a semelhança entre as observações existentes da classe minoritária. Para cada observação x_i da classe minoritária é gerada uma observação sintética de acordo com:

$$x_{sintético} = x_i + (y_i - x_i) \times gap$$

onde y_i é um dos k-vizinhos mais próximos de x_i e gap é um valor entre 0 e 1. Conclui-se desta expressão, que o resultado gerado é um ponto ao longo de uma recta, unindo

o ponto x_i com cada um dos k-vizinhos mais próximos da classe minoritária. Após a realização do balanceamento de classes obtiveram-se 7772 observações na amostra final, contendo 3886 clientes em incumprimento e 3886 clientes regulares.

Capítulo 4

Técnicas de previsão do incumprimento

Neste estudo foram implementados modelos de previsão do incumprimento baseados na regressão logística, em árvores de decisão e nos classificadores múltiplos *Bagging*, *Boosting* e *Voting*.

4.1 Regressão Logística

A regressão logística é apropriada nas situações em que a variável dependente é binária (Hosmer e Lemeshow, 2000). A partir de um conjunto de variáveis independentes, $x_1, \ldots, x_k$, este método estima a probabilidade de ocorrer um determinado evento. Neste estudo, foi atribuído à variável resposta o valor um para indicar incumprimento e o valor zero no caso contrário. A função utilizada na regressão logística para estimar a probabilidade de uma determinada realização i da variável resposta ser o "sucesso", $P(Y_i = 1) = \hat{p}_i$, é a função logística que possui a seguinte forma:

$$\hat{p}_i = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_k x_{ki}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_k x_{ki}}}$$

A equação de regressão produz:

$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_k x_{ki} = logit(\hat{p}_i)$$

Os coeficientes $\hat{\beta}_i$ são estimados pelo método da máxima verossimilhança. Neste método, os coeficientes estimados maximizam a probabilidade de obter as realizações da variável dependente da amostra em estudo (Hosmer e Lemeshow, 2000).

4.1.1 Teste de adequação do modelo

Após o ajuste do modelo de regressão logística é necessário avaliar a significância do modelo ajustado, tal como a significância dos coeficientes de regressão. A significância do modelo ajustado é obtida pelo teste das seguintes hipóteses:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \quad vs \quad H_1 : \exists_i : \beta_i \neq 0 \quad (i = 1, ..., k)$$

É possível prever a probabilidade do evento, a partir das variáveis independentes do modelo, apenas quando o modelo ajustado é estatisticamente significativo, condição que é satisfeita quando a hipótese alternativa se verifica. Para testar a significância do modelo, utiliza-se uma estatística de teste que compara a verossimilhança de um modelo que contenha apenas a constante (ou seja, nenhuma das variáveis independentes tem poder de previsão) com a verossimilhança do modelo que contém as variáveis independentes.

A estatística para testar a significância do modelo de regressão logística é dada por:

$$G^2 = -2\ln\left(\frac{L_0}{L_C}\right) \quad \overset{a}{\sim} \quad \chi^2_k$$

onde L_0 é a função verosimilhança para o modelo que contém somente a constante e L_C é a função verosimilhança para o modelo completo. Rejeita-se a hipótese nula H_0 se o valor-p do G^2 observado for inferior à dimensão do teste α.

É importante referir, que o facto do modelo ajustado ser estatisticamente significativo, permite apenas afirmar que pelo menos uma das variáveis independentes do modelo influencia significativamente a variável dependente.

4.1.2 Significância dos coeficientes do modelo

Quando se pretende identificar qual ou quais as variáveis independentes que influenciam significativamente a variável resposta é usual recorrer-se ao Teste de Wald . O objectivo é testar se um determinado coeficiente é nulo, condicionado pelos valores estimados dos outros coeficientes:

$$H_0 : \beta_i = 0 \mid \beta_0, \beta_1, \beta_{i-1}, \beta_{i+1}, \beta_k \ \ vs \ \ H_1 : \beta_i \neq 0 \mid \beta_0, \beta_1, \beta_{i-1}, \beta_{i+1}, \beta_k \ \ (i = 1, ..., k)$$

A estatística de teste possui a seguinte expressão:

$$T_{Wald_i} = \frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)} \quad \overset{a}{\sim} \ N(0, 1)$$

onde, $\hat{\beta}_i$ é o estimador de β_i e $\hat{SE}(\hat{\beta}_i)= \sqrt{\hat{\sigma}^2(\hat{\beta}_i)}$ é o estimador do desvio-padrão de $\hat{\beta}_i$, calculado através da função de Informação de Fisher $\mathbf{I}(\hat{\boldsymbol{\beta}})=\mathbf{X'VX}$ (Marôco, 2011). A matriz de variância-covariância dos parâmetros do modelo é $\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})$ em que o i-ésimo elemento da diagonal principal é a estimativa de $\hat{\sigma}^2(\hat{\beta}_i)$. A hipótese nula é rejeitada para cada um dos testes aos β_i quando o respectivo valor-p for inferior à dimensão do teste α.

4.1.3 Resultados - Regressão Logística

O modelo final de previsão do incumprimento, obtido pela regressão logística, é apresentado na Tabela 4.1. Nesta tabela constam as estimativas dos coeficientes de regressão e a sua significância no modelo.

Atendendo ao valor-p associado às estimativas dos coeficientes $\hat{\beta}_5$, $\hat{\beta}_{10}$, $\hat{\beta}_{18}$ conclui-

	$\hat{\beta}_i$	valor-p
I1	-0.1632	0.000
I2	-2.5313	0.000
I3	-1.6411	0.000
I4	-2.8913	0.000
I5	-0.0002	0.507
I6	-1.0302	0.000
I7	1.6431	0.000
I8	-1.7452	0.000
I9	8.0995	0.000
I10	-0.0171	0.648
I11	-2.2902	0.001
I12	-0.0009	0.031
I13	-1.7741	0.000
I14	-0.034	0.016
I15	0.7046	0.000
I16	2.4891	0.000
I17	5.7267	0.000
I18	0.02	0.663
Constante	-0.8173	0.000

Tabela 4.1: Modelo de previsão do incumprimento dado pela regressão logística.

se que estes coeficientes não são significativos no modelo, para um nível de significância $\alpha = 0.05$. No entanto, como o principal objectivo deste estudo é a comparação do desempenho de diversas técnicas utilizadas na previsão do incumprimento, optou-se por manter todas as variáveis no modelo. O rácio de verossimilhança possui um valor de 1442.86 e um valor-p inferior a 0.001. Portanto, o modelo é estatisticamente significativo. Analisando os sinais dos coeficientes $\hat{\beta}_j, j = 1, \ldots, 18$, na Tabela 4.1, conclui-se que, com a excepção dos rácios I1, I2 e I7, praticamente todos os coeficientes estatisticamente significativos estão de acordo com a teoria económica-financeira. Eliminando os rácios I1, I2 e I7 do modelo observa-se uma degradação da precisão na classificação dos clientes quanto ao incumprimento. Deste modo, optou-se por manter no modelo todas as variáveis.

A classificação dos clientes quanto ao incumprimento é realizada com base na pontuação ("score") dada pela regressão logística. Como os dados estão balanceados, utilizou-se um ponto de corte de 0.5. Assim, os clientes com pontuação inferior a 0.5 são classificados como regulares, enquanto os clientes com pontuação superior a

0.5 são classificados em situação de incumprimento.

4.2 Árvores de Decisão

As árvores de decisão são modelos não-paramétricos que podem ser utilizados em problemas de classificação ou de regressão [1] (Breiman et al., 1984). Estas estruturas consistem numa sequência de regras que divide os dados em subconjuntos mutuamente exclusivos. Estas regras são representadas pelos testes, realizados nos diversos nós, sobre os atributos (por exemplo, rácios económico-financeiros) das observações. A divisão dos dados é realizada até que cada subconjunto resultante das sucessivas partições contenha uma clara maioria de uma das classes, não se justificando posteriores divisões. A classificação final, ou seja, a classificação do crédito em situação de incumprimento ou em situação regular para uma observação é determinada pelo percurso da raíz ao nó terminal, ditado pelos diversos testes presentes ao longo da árvore.

Existem diversas vantagens na utilização das árvores de decisão: estes modelos não assumem nenhuma distribuição particular para os dados; a estrutura da árvore é independente da escala das variáveis; os modelos apresentam um elevado grau de interpretabilidade; a construção dos modelos é computacionalmente eficiente; o algoritmo é munido de um mecanismo de selecção de atributos, sendo robusto à presença de *outliers* e a atributos redundantes ou irrelevantes. O método apresenta como desvantagem a instabilidade, dito de outra forma, pequenas perturbações do conjunto de treino podem provocar grandes alterações no modelo estimado (Roe et al., 2005).

4.2.1 Algoritmo *REPTree*

Existem diversos algoritmos de construção de árvores de decisão. Neste estudo, usou-se o algoritmo *REPTree* que utiliza o ganho de informação (Mitchell, 1997) na definição do teste a ser executado em cada nó de decisão. O ganho de informação

[1] Problemas que admitem classes contínuas.

representa o incremento de informação produzido pela partição do conjunto de treino de acordo com seu candidato à partição (Mitchell, 1997). O algoritmo particiona o conjunto de dados levando em consideração a variável que produz o melhor ganho de informação. Deste modo, pode-se dizer que o atributo que melhor classificar os dados deve ser escolhido como um nó da árvore.

Com o objectivo de evitar o sobre-ajustamento [2] aos dados de estimação dos parâmetros do modelo e reduzir a dimensão da árvore, o algoritmo efectua a "poda" da árvore com recurso à técnica *reduced-error pruning* (REP), a qual minimiza o erro num conjunto de validação independente do conjunto de treino (Witten e Frank, 1999). O procedimento de efectuar a poda com redução do erro, examina cada nó interno da árvore de decisão e substitui-o pelo melhor nó terminal possível caso o número de erros de classificação não aumente, ignorando os nós com pouca relevância para a classificação (Quinlan, 1986).

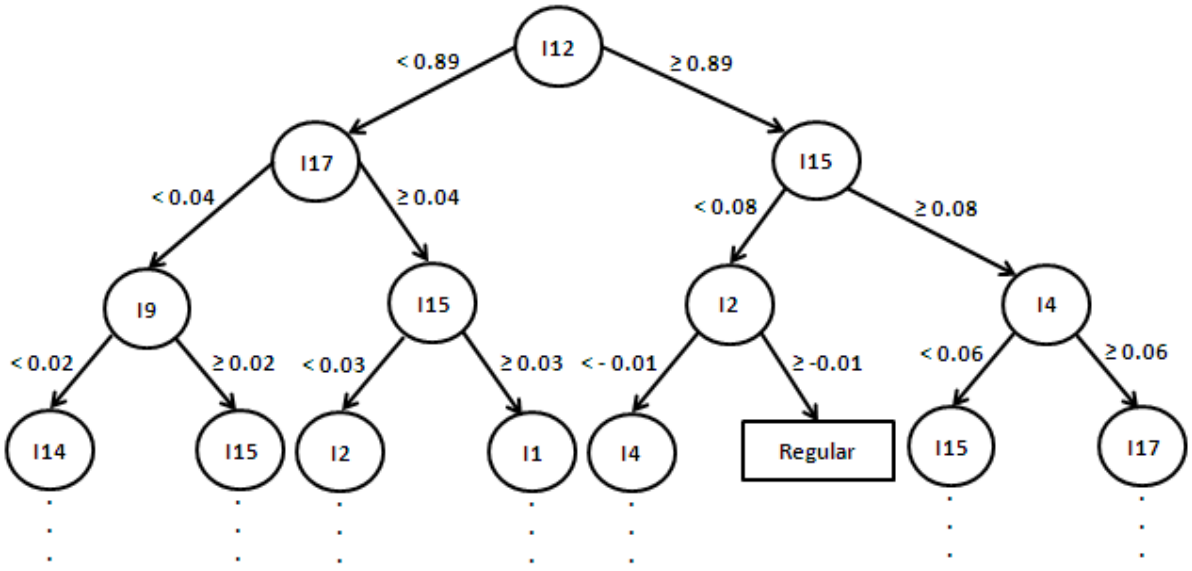

Figura 4.1: Representação parcial do modelo de previsão do incumprimento obtido pelo algoritmo de árvore de decisão.

Na Figura 4.1 é apresentado parte do modelo de previsão do incumprimento

[2] Também designado como *over-fitting*, ocorre quando o classificador tende a se adaptar a detalhes específicos da amostra de treino, o que pode causar em dados futuros uma redução da taxa de acerto.

gerado pelo algoritmo de árvore de decisão. Por exemplo, um crédito caracterizado por um valor do rácio I12 superior a 0.89, um valor do rácio I15 inferior a 0.08, e um valor do rácio I2 superior a -0.01, é classificado pela árvore de decisão como regular.

4.3 Classificadores múltiplos

Os classificadores múltiplos, também designados por *ensembles*, representam métodos onde diversos classificadores são treinados por forma a solucionar o mesmo problema. A aplicação destes métodos apenas é vantajosa se os classificadores base possuirem um bom desempenho individual mas também um comportamento diverso em relação aos demais (Krogh e Vedelsby, 1995). Dito de outra forma, a diversidade ocorre quando os classificadores base apresentam erros descorrelacionados no espaço das observações, tornando desta maneira eficaz a combinação dos diferentes classificadores. No presente estudo foram utilizados dois classificadores múltiplos homogéneos, *Bagging* e *Boosting*, e um classificador múltiplo heterogéneo, *Voting*.

4.3.1 Classificadores múltiplos homogéneos

Os classificadores múltiplos homogéneos combinam modelos gerados por um único algoritmo, manipulando o conjunto de treino por forma a gerar múltiplas hipóteses (Gama et al., 2011). Esta técnica é adequada especialmente para algoritmos de classificação instáveis, nos quais os resultados dos classificadores possuem grandes alterações em resposta a pequenas mudanças nos dados de estimação dos parâmetros do modelo. Deste modo, as árvores de decisão são uma boa escolha para classificador base dos classificadores múltiplos homogéneos.

4.3.1.1 *Bagging*

O método *Bagging - Bootstrap Aggregating* (Breiman, 1996), constrói os classificadores com base em réplicas do conjunto de treino obtidas através de amostragem com

reposição. Do conjunto de treino obtém-se réplicas que possuem o mesmo tamanho que os dados originais, não constando algumas observações da amostra original e com a possibilidade de outras surgirem repetidas vezes. Existe desta forma, uma replicação e ausência de certos exemplos, criando classificadores diferentes devido à variação de exemplos nas amostras. Neste método, a variabilidade aleatória dos classificadores individuais é reduzida devido ao voto maioritário de diferentes hipóteses (Bauer e Kohavi, 1999). Por esta razão, através desta técnica obtém-se uma melhoria nas árvores de decisão, dado que este é um algoritmo instável. O pseudo-código deste método encontra-se no anexo A.2.

4.3.1.2 *Boosting*

O classificador múltiplo *Boosting* gera vários classificadores sequencialmente. Em cada iteração o algoritmo altera a distribuição do conjunto de treino em função das classificações anteriores (Witten et al., 2011). Neste estudo usou-se um algoritmo de *boosting* designado por *AdaBoost* (*Adaptive Boosting*) e que foi proposto por Freund e Schapire (1996).

O algoritmo resume-se nas seguintes etapas: inicialmente atribui a todos as observações de treino o peso $1/n$, em que n é a quantidade de observações do conjunto de treino; o classificador é então treinado de acordo com a distribuição de pesos na i-ésima iteração D_i e posteriormente calcula-se o erro e_i nessa iteração; constrói-se uma nova distribuição de pesos D_{i+1}, diminuindo os pesos dos que foram classificados correctamente (multiplica-se por $e_i/(1-e_i)$) e aumentando o peso das observações classificadas erroneamente; normaliza-se o peso de todas as observações, um novo treino é realizado com a nova distribuição de pesos; os erros e pesos são actualizados e o processo repetido N vezes; por fim, obtém-se o classificador final através da agregação dos classificadores aprendidos em cada iteração pela votação pesada (Freund e Schapire, 1999). O pseudo-código deste algoritmo encontra-se no anexo A.3.

4.3.2 Classificadores múltiplos heterogéneos

Os classificadores múltiplos heterogéneos utilizam diferentes algoritmos como classificadores base.

4.3.2.1 *Voting*

O classificador múltiplo *Voting* (Kittler, 1998), combina classificadores distintos através de um esquema de votação. Um dos esquemas de votação mais simples é o de votação maioritária. Este exige que cada classificador base apresente como saída um voto à classe que considere ser a mais provável para um dado exemplo. Desta forma, realiza-se uma contagem do número de votos por classe para todos os classificadores. Por fim, escolhe-se a classe com maior número de votos como previsão final para a observação em estudo. Algumas variações dessa ideia originaram os diversos métodos de votação, como por exemplo: média das probabilidades, produto das probabilidades, probabilidade mínima e probabilidade máxima. Portanto, num problema com m classificadores e j classes, têm-se as seguintes fórmulas de cálculo:

- Média das Probabilidades: $S_j = \sum_{k=1}^{m} \frac{p_{kj}}{m}$
- Produto das Probabilidades: $S_j = \Pi_{k=1}^{m} \, p_{kj}$
- Probabilidade Mínima: $S_j = min_k \, p_{kj}$
- Probabilidade Máxima: $S_j = max_k \, p_{kj}$

Por fim, a observação deve ser classificada na classe que maximiza S_j. Neste estudo, os modelos base para o classificador *Voting* foram a regressão logística e a árvore de decisão.

Capítulo 5

Avaliação do poder de previsão

A qualidade das previsões produzidas pelos diferentes modelos é avaliada através da área sob a curva ROC (AUC) e das taxas de erro de classificação. A capacidade preditiva dos modelos é avaliada "dentro da amostra" e "fora da amostra". A avaliação "dentro da amostra" é realizada nos dados usados na estimação dos modelos. Em geral, esta avaliação é demasiado optimista uma vez que os modelos tendem a sobreajustar os dados usados na estimação. A avaliação "fora da amostra" é realizada através de uma validação cruzada com 10-*folds*. Esta técnica divide os dados em 10 subconjuntos, utilizando um dos subconjuntos para teste e realizando o treino com os demais; este procedimento é repetido 10 vezes alternando o conjunto de teste. O erro de classificação é dado pela média dos erros calculados em cada uma das 10 iterações.

5.1 Curva ROC

A curva ROC (*Receiver Operating Characteristics*) é um método eficiente na análise do desempenho de algoritmos de classificação e que é particularmente útil quando os dados possuem custos de classificação diferentes por classe. Esta técnica consiste num gráfico de pares (x, y) num plano, no qual o eixo das ordenadas representa a sensibilidade do modelo, ou seja, o quão eficaz é o modelo em prever *verdadei-*

ros positivos (i.e., créditos em incumprimento), e o eixo das abcissas representa o complementar da *especificidade*. A especificidade representa a capacidade do modelo não errar na identificação de *verdadeiros negativos* (i.e., créditos regulares). O ponto (0,1) no plano representa o classificador perfeito, ou seja, no qual todos os exemplos positivos são classificados correctamente e nenhum exemplo negativo é classificado como positivo. A curva ROC permite estudar a variação da sensibilidade e especificidade para diferentes pontos de corte. O valor do ponto de corte, ou seja, o valor acima do qual o cliente é classificado como em situação de incumprimento (positivo) e abaixo do qual é classificado como regular (negativo) é definido pelas instituições financeiras.

A área abaixo da curva ROC avalia a capacidade do modelo para discriminar indivíduos com factor de interesse em estudo relativamente aqueles que não têm o factor de interesse. Quanto maior esta área, melhor é o desempenho médio do classificador. O valor da área abaixo da curva ROC igual a 1 indica que se tem um modelo perfeito, um valor de cerca de 0.5 caracteriza um modelo aleatório possuindo uma fraca capacidade de discriminação.

5.2 Análise dos erros

A avaliação do desempenho de um classificador $\hat{c}$ também pode ser realizada através da sua taxa de erro de classificação. Esta taxa é dada por:

$$erro(\hat{c}) = \frac{1}{n} \sum_{i=1}^{n} I(\hat{c}(x_i) \neq y_i)$$

onde n é o número de observações nos dados, $I(z) = 1$ se a condição z é verdadeira e $I(z) = 0$ caso contrário. Esta taxa obtém-se pela comparação da classe conhecida de x_i, com a classe prevista.

Numa previsão, e para um determinado valor de corte, pode cometer-se dois tipos de erro: o erro tipo I e o erro tipo II. No caso da previsão de incumprimentos

de crédito, o erro tipo I consiste em classificar como regular clientes que virão a incumprir e neste caso a entidade está exposta ao risco de crédito. Já o erro tipo II, corresponde a classificar em situação de incumprimento os contratos que não possuem esta característica. Quando este erro é elevado por um longo período de tempo, haverá perdas na concessão de crédito, risco de perda de quota no mercado e de quebra nos lucros. É importante realçar que no processo de gestão de risco de crédito, o erro tipo II é mais aceitável por ser conservador, dado que o erro de aprovar uma operação que se tornará problemática (erro tipo I) é considerado mais grave que a recusa de uma operação que seria um bom negócio para a instituição (erro tipo II).

Capítulo 6

Resultados

Foi investigado o desempenho na classificação de incumprimentos de modelos baseados na regressão logística, na árvore de decisão e nos classificadores múltiplos *Bagging*, *Boosting* e *Voting*. O modelo base nos classificadores *Bagging* e *Boosting* foi a árvore de decisão. Os modelos base para o classificador *Voting* foram a regressão logística e a árvore de decisão. O desempenho dos métodos foi avaliado através da área sob a curva ROC e das taxas de erro de classificação. O número de iterações nos classificadores *Bagging* e *Boosting* foi obtido através da maximização da área abaixo da curva ROC dada pela validação cruzada.

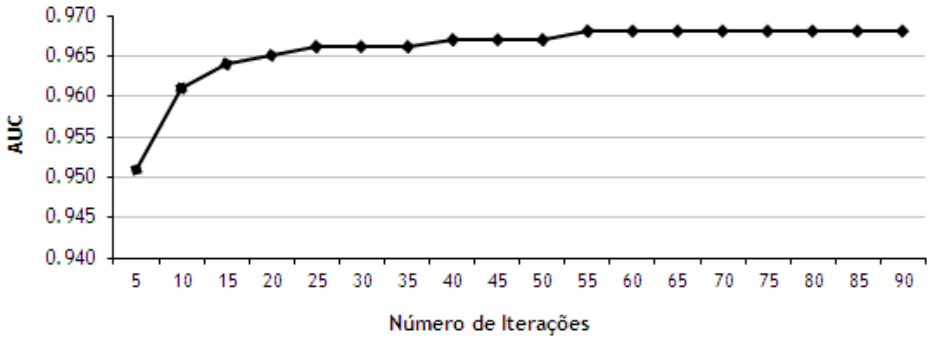

Figura 6.1: Área sob a curva ROC em função do número de árvores de decisão no classificador múltiplo *Bagging*.

A Figura 6.1 apresenta a AUC em função do número de árvores de decisão no classificador múltiplo *Bagging*. À medida que aumenta o número de árvores,

aumenta a AUC e, logo, a precisão do classificador múltiplo. No entanto, o impacto marginal de cada árvore adicionada é decrescente. Após serem adicionadas cerca de 55 árvores ao classificador múltiplo não se observam melhorias significativas na AUC.

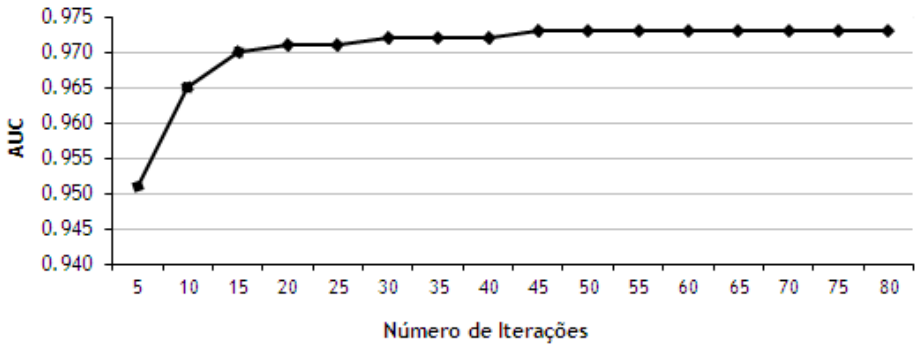

Figura 6.2: Área sob a curva ROC em função do número de árvores de decisão no classificador múltiplo *Boosting*.

A Figura 6.2 apresenta a AUC em função do número de árvores de decisão no classificador múltiplo *Boosting*. O comportamento da AUC em função do número de membros do classificador múltiplo é semelhante. No entanto, a AUC estabiliza quando são adicionadas cerca de 45 árvores.

Método de Votação	**AUC**
Média das Probabilidades	0.897
Produto das Probabilidades	0.895
Voto Maioritário	0.726
Probabilidade Mínima	0.895
Probabilidade Máxima	0.894

Tabela 6.1: Área sob a curva ROC do classificador múltiplo *Voting* para os diferentes métodos de votação.

A Tabela 6.1 apresenta a área sob a curva ROC do classificador múltiplo *Voting* para os diferentes métodos de votação. Com excepção do método de voto maioritário, todos os métodos de votação têm um desempenho semelhante.

A Tabela 6.2 compara a AUC de todos os métodos considerados. Entre os classificadores simples, a árvore de decisão teve um desempenho melhor do que a regressão

Algoritmo	**AUC**
Regressão logística	0.768
Árvore de decisão	0.894
Bagging de árvores de decisão	0.968
Boosting de árvores de decisão	0.973
Voting (regressão logística e árvore de decisão)	0.897

Tabela 6.2: Área sob a curva ROC dada pelos diferentes tipos de classificador.

logística. Por outro lado, o classificador múltiplo *Voting* teve um desempenho marginalmente superior ao da árvore de decisão. Os melhores resultados são dados pelos classificadores múltiplos homogéneos *Bagging* e *Boosting*. Em particular, o classificador *Boosting* teve um desempenho ligeiramente superior ao classificador *Bagging*. As áreas sob a curva ROC superiores a 0.95 indicam uma excelente capacidade destes modelos para distinguir empresas em situação de incumprimento das empresas regulares. A superioridade dos classificadores múltiplos homogéneos na previsão do incumprimento deve-se à capacidade destes em aproximar funções mais complexas, e portanto construir fronteiras de decisão que proporcionam classificadores com maior poder preditivo.

Classificador	Taxa de Erro	
	dentro da amostra	fora da amostra
Regressão logística	28.9%	28.9%
Árvore de decisão	9.8%	14.9%
Bagging de árvores de decisão	4.1%	9.6%
Boosting de árvores de decisão	0.8%	8.0%
Voting (regressão logística e árvore de decisão)	10.4%	15.1%

Tabela 6.3: Taxas de erro dos classificadores dentro da amostra e fora da amostra.

Na Tabela 6.3 são apresentadas as percentagens de erro de classificação de cada método, dentro da amostra e fora da amostra. Para calcular estes erros, os créditos foram classificados como regulares se a pontuação dada pelo classificador foi inferior a 0.5, e foram classificados como em incumprimento se a pontuação foi superior a 0.5. Constata-se que para todos os classificadores excepto a regressão logística o erro fora da amostra é superior ao erro dentro da amostra. Isto indica que os classificadores baseados em árvores de decisão tendem a sobre-ajustar os dados usados na estimação

dos modelos. Atendendo aos erros fora da amostra (os mais relevantes para efeitos de previsão) verifica-se mais uma vez que os classificadores múltiplos homogéneos apresentam o melhor desempenho na classificação dos créditos. É interessante notar que o desempenho do classificador *Voting* não é superior ao do classificador baseado numa árvore de decisão individual.

	Valores Previstos			
Valores Observados	Incumprimento	Regular	Total	% de Acerto
Incumprimento	3605	281	3886	92.77%
Regular	342	3544	3886	91.20%
Total	3947	3825	7772	91.98%

Tabela 6.4: Matriz de classificação dada pelo classificador *Boosting*.

Conforme foi referido, os erros de previsão do tipo I e do tipo II não têm o mesmo custo para as instituições financeiras. No entanto, em alguns casos as instituições financeiras optarão por conceder o crédito mesmo que este apresente características de créditos em situação de incumprimento. Isto ocorre pelo facto da concessão do empréstimo poder vir a ser benéfico para a instituição. Neste casos, determina-se prazos e montantes de empréstimo menores e taxas de juros mais elevada. A Tabela 6.4 apresenta o número de erros do tipo I e do tipo II cometidos pelo melhor classificador: o *Boosting* de árvores de decisão. Este classificador apresentou uma menor taxa de erro do tipo I relativamente a taxa de erro do tipo II. Constata-se também, que a percentagem de acerto obtida através deste modelo foi elevada (91.98%). A percentagem de empresas previstas que entrariam em situação de incumprimento e que, de facto, entraram em incumprimento foi de 92.77%. Isto indica que 92.77% das ocorrências de incumprimento na amostra de empresas foram correctamente previstas. Estes resultados confirmam a excelente capacidade de previsão do modelo.

Capítulo 7

Considerações finais

Neste estudo foram implementados diferentes classificadores múltiplos para previsão do incumprimento no crédito a empresas. Os classificadores múltiplos considerados foram o *Bagging*, o *Boosting* e o *Voting*. Utilizando informação extraída de uma base de dados de uma instituição bancária portuguesa, este estudo sugere que os classificadores múltiplos apresentam melhor capacidade de previsão do incumprimento do que as técnicas tradicionais, como a regressão logística e as árvores de decisão. Em particular, a técnica de *Boosting* de árvores de decisão obteve o melhor desempenho, seguida da técnica de *Bagging* de árvores de decisão.

As técnicas apresentadas neste trabalho demonstraram ser ferramentas de grande valor para os analistas de crédito a empresas. Utilizando os rácios económico-financeiros, os analistas têm condições de diagnosticar os novos clientes quanto à concessão de crédito ou não. A experiência profissional do analista de crédito, aliada às técnicas de classificação utilizadas neste estudo, são instrumentos que podem ajudar na tarefa de tomada de decisão. É importante destacar que o principal objectivo dos modelos de previsão de incumprimento não é ditar a decisão final sobre a concessão de crédito, mas sim, fornecer aos analistas informações que os auxiliem a tomar decisões mais direccionadas e correctas.

Apêndice A

Anexos

A.1 Pseudo-Código SMOTE

Entrada Algoritmo SMOTE(t,n,k):

- Número de exemplos da classe minoritária t
- Aumento da classe minoritária em n%
- Número de k-vizinhos mais próximos

Saída: $(n/100)\times$t exemplos sintéticos da classe minoritária

1 Se n é menor que 100%, escolhe-se uma amostra aleatória da classe minoritária para que sobre esta seja aplicada o método

2 if $n < 100$

3 então escolher uma amostra aleatória da classe minoritária

4 $t = (n/100) \times t$

5 n=100

6 endif

7 n=(int)(n/100)

8 k= número de vizinhos mais próximos

9 NumAtrib= número de atributos

10 Amostra[ ][ ]: vector para os exemplos originais da classe minoritária

11 NovoÍndice: contador das amostras sintéticas geradas inicializado a 0

12 Sintético[ ][ ]: vector para as amostras sintéticas

Calcular os k-vizinhos mais próximos apenas para cada exemplo da classe minoritária

13 for i $\leftarrow$ 1 até t

14 Calcular os k-vizinhos mais próximos para i e guardar o índice no vector narray

15 População(n,i,narray): função que gera a amostra sintética

16 end for

População(n,i,narray)

17 while $n \neq 0$

18 Escolher um número aleatório entre 1 e k (nn). Este passo escolhe um dos k-vizinhos mais próximos de i.

19 for Atrib $\leftarrow$ 1 to NumAtrib

20 Calcular: dif=Amostra[narray[nn]][Atrib]-Amostra[i][Atrib]

21 Calcular: gap= número aleatório entre 0 e 1

22 Sintético[NovoÍndice][Atrib]= Amostra[i][Atrib]+ gap$\times$dif

23 end for

24 NovoÍndice ++

25 n=n-1

26 end while

27 return (Fim da População)

A.2 Pseudo-Código *Bagging*

Entrada Algoritmo *Bagging*:

- Classificador base **c**
- Conjunto de treino $\mathbf{D} = \{(x_i, y_i), i = 1, ..., n\}$
- Número de Iterações N
- Conjunto de teste contendo t exemplos $\mathbf{T} = \{(x_j, ?), j = 1, ..., t\}$

1 Aprendizagem

2 for $l = 1$ **to** N **do**

3 $\mathbf{D}^* \leftarrow$ amostra com reposição de **D**

4 $\hat{c}_l \leftarrow c(\mathbf{D}^*)$

5 end for

6 Classificação

7 for $j = 1$ **to** t **do**

8 $\hat{y}_j = argmax_{y \in Y} \sum_{l=1}^{N} I(\hat{c}_l(x_j \in \mathbf{T}) = y)$

9 end for

10 Retorna: Vector de previsões $\hat{y}$

Saída Algoritmo *Bagging*: Previsões para o conjunto de teste

A função I(.) devolve 1 se a condição for verdadeira e 0 caso contrário.

A.3 Pseudo-Código *AdaBoost*

Entrada Algoritmo:

- Classificador base **c**
- Conjunto de treino **D** $=\{(x_i, y_i), i = 1, ..., n\}$
- Número de Iterações N
- Conjunto de teste com t exemplos $\mathbf{T} = \{(x_j, ?), j = 1, ..., t\}$

1 Treino

2 for $x_i \in$ **D do**

3 $w(x_i) \leftarrow 1/\mathrm{n}$;

4 end for

5 for $\mathrm{l} = 1$ **to** N **do**

6 **for** $x_i \in$ **D do**

7 $p_l(x_i) \leftarrow w_l(x_i)/\sum_i w_l(x_i)$;

8 **end for**

9 **Invocação do Algoritmo de Aprendizagem**

10 $c_l^* \leftarrow c(p_l)$;

11 **Calcular o Erro**

12 $e_l = \sum_i p_l(x_i)[c_l^*(x_i) \neq y_i]$;

13 $\alpha_l \leftarrow log(\frac{1-e_l}{e_l})$;

14 **for** $x_i \in$ **D do**

15 $w_{l+1}(x_i) := w_l(x_i)exp[\alpha_l I(c_l^*(x_i) \neq y_i)]$;

16 **end for**

17 end for

18 Fase de Teste

19 do $\mathrm{j} = 1$ **to** t **do**

20 $\hat{y_j} = \arg\, max_{y \in Y} \sum_{l=1}^{N} \alpha_l [c_l^*(x_j \in \mathbf{T}) = y]$;

21 end for

22 Retorna: Vector de previsões $\hat{y}$;

Saída Algoritmo: Previsões para o conjunto de teste

Na linha 13, o termo α_l representa o peso do classificador l. Note-se ainda, que na linha 15, os pesos dos exemplos classificados incorrectamente são ajustados (Gama et al., 2011).

Bibliografia

[1] Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance, 23(4), 589-609.

[2] Baesens, B., Gestel, T. V., Viaene, S., Stepanova, M., Suykens, J. e Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54, 627-635.

[3] Bauer, E. e Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, Boosting, and Variants. Machine Learning, 36, 105-142.

[4] Beaver, W. H. (1966). Financial Ratios as Predictors of Failure. Journal of Accounting Research, 4, 71-111.

[5] Breiman, L., Friedman, J. H., Olshen, R. A. e Stone, C. J. (1984). Classification and Regression Trees. Belmont, CA: Wadsworth.

[6] Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2): 123-140.

[7] Carvalho, C. N. e Magalhães, G. (2002). Análise Económico-Financeira de Empresas. 1ª Ed. Lisboa: Universidade Católica Editora.

[8] Chawla, N. V., Bowyer, K. W., Hall, L. O. e Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

[9] Freund, Y. e Schapire, R. E. (1996). Experiments with a New Boosting Algorithm, In International Conference on Machine Learning, 148-156.

[10] Freund, Y. e Schapire, R. E. (1999). A Short Introduction to Boosting. Japonese Society for Artificial Intelligence, 5, 771-780.

[11] Frydman, H., Altman, E. I. e Kao, D. L. (1985). Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress. Journal of Finance 40(1), 269-291.

[12] Gama, J., Faceli, K., Lorena, A. C. e Carvalho, A. (2011). Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina. 1ª Ed. Rio de Janeiro: LTC.

[13] Henley, W. E. e Hand, D. J. (1996). A k-nearest neighbour classifier for assessing consumer credit risk. The Statistician, 45(1), 77-95.

[14] Hosmer, D.W. e Lemeshow, S. (2000). Applied Logistic Regression. 2ª Ed. Usa: Wiley & Sons.

[15] Jensen, H. L. (1992). Using neural networks for credit scoring. Managerial Finance, 18(6), 15-26.

[16] Kittler, J. (1998). Combining classifiers: A theoretical framework. Pattern Analysis and Applications, 1(1), 18-27.

[17] Krogh, A. e Vedelsby, J. (1995). Neural Network Ensembles, Cross Validation, and Active Learning. Advances in Neural Information Processing Systems, 231-238.

[18] Marôco, J. (2011). Análise Estatística com o SPSS Statistics. 5ª Ed. Pero Pinheiro: ReportNumber.

[19] Mitchell, T. (1997). Machine Learning.

[20] Quinlan, J. R. (1986). Simplifying Decision Trees.

[21] Reichert, A. K., Cho, C. C. e Wagner, G. M. (1983). An Examination of the Conceptual Issues Involved in Developing Credit-scoring Models. Journal of Business & Economic Statistics, 1(2), 101-114.

[22] Roe, B. P., Yang, H. J. e Zhu, J. (2005). Boosted Decision Trees, A Powerful Event Classifier.

[23] West, D., Dellana, S. e Qian, J. (2005). Neural network ensemble strategies for financial decision applications. Computers and Operations Research, 32, 2543-2559.

[24] Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. Journal of Financial and Quantitative Analysis, 15, 757-770.

[25] Witten, I. e Frank, E. (1999). Reduced-error pruning with significance tests.

[26] Witten, I., Frank, E. e Hall, M. (2011). Data Mining Practical Machine Learning Tools and Techniques. 3ª Ed. USA: Elsevier.