2
Banco de Dados

Grandes volumes de dados são difíceis, e muitas vezes, impossíveis de serem obtidos. Além disso, os dados históricos disponíveis nem sempre são bem representativos da realidade e podem comprometer resultados de modelos que serão usados a posteriori. Não obstante, a implementação de modelos aplicáveis em investimentos, deve considerar que sistemas que utilizem estes modelos sejam capazes de predizer o futuro em tempo real e através de dados históricos confiáveis.

No presente estudo, utilizou-se um banco de dados composto por 418 das 500 empresas que compõem o índice S&P500 da Standard & Poor's. O S&P500, é um índice que consiste em ações de 500 companhias líderes nos setores mais importantes da economia norte-americana e escolhidas por tamanho de mercado, liquidez e representação de grupo industrial por um comitê da S&P. Este índice está entre os indicadores mais importantes do mercado de capitais dos Estados Unidos, com o seu desempenho refletido na Bolsa de Valores de Nova York.

Este banco de dados não contém instituições financeiras e apesar das informações nele contidas serem de domínio público, foi concatenado e organizado em planilha eletrônica pela Thomson Financial Services[1]. Este inicialmente contém 418 empresas (observações), com seus respectivos *ratings* em 2005 e 19 atributos referentes a 2004, a saber:

1. Setor;
2. Vendas;
3. Receitas;
4. Lucro antes de Impostos e Juros *(EBIT)*;
5. Depreciação;
6. Receitas Financeiras;

[1] www.thomson.com

PUC-Rio - Certificação Digital Nº 0321206/CA

7. Despesas Financeiras;
8. Lucro Líquido (conforme) Publicado;
9. Lucro Líquido;
10. Ativo Total;
11. Ativo Corrente;
12. Passivo corrente;
13. Passivo Total;
14. Endividamento Total;
15. Disponibilidade;
16. Endividamento Líquido;
17. Liquidez Corrente;
18. Valor de Mercado;
19. Despesas de Amortização.

PUC-Rio - Certificação Digital Nº 0321206/CA

Durante a análise e preparação da Base de Dados, realizou-se um criterioso tratamento de dados faltantes, descrito na seção 2.1 e a posterior formatação do banco de dados, descrito na seção 2.2.

2.1 Tratamento de dados faltantes

Lidar com dados faltantes é particularmente desafiador em um contexto onde os dados disponíveis são preciosos. A literatura na análise de dados parcialmente faltantes é comparativamente recente, tendo iniciado em meados dos anos 60 e se estendendo ao longo das últimas décadas. É interessante notar que até dados faltantes podem obedecer a certos padrões (Little, 2002), por exemplo, ocorrer de forma univariada, multivariada, ou monótona, dentre outros.

Uma proposta para resolver este problema é simplesmente eliminar as observações que tenham dados faltantes. Outra alternativa é assumir que o dado faltante é igual à um dado de um período financeiro imediatamente anterior (Deboeck, 1994). Fundamentalmente, esta decisão depende da qualidade e do volume existente dos dados e dos métodos e modelos que serão aplicados sobre

estes dados. Métodos pouco robustos e pouco tolerantes a falhas tendem a serem mais sensíveis a dados faltantes que outros que não possuam estas características.

De maneira geral, podemos agrupar os métodos de tratamento de dados faltantes em quatro categorias, não mutuamente exclusivas, a saber (Deboeck, 1994):

a. Tentar conseguir dados de outras fontes;
b. Interpolar os valores ou usar valores médios;
c. Preencher com valores que minimizem a alteração de padrões;
d. Eliminar a observação.

Idealmente a primeira opção é a melhor, porém nem sempre isso é possível. Devido aos algoritmos de treinamento de Redes Neurais serem robustos, *outliers* em pequenas quantidades ou dados ruidosos tendem a não comprometer seus resultados. Funções de ativação tais como sigmóide e tangente hiperbólica naturalmente saturam um dado *outlier*, de forma que procedimentos que identificam e retiram *outliers* podem ser úteis, sobretudo, para acelerar o aprendizado das Redes Neurais. Porém há a desvantagem do risco de se omitir alguma informação importante contida em um suposto *outlier*, uma vez que a sua identificação é subjetiva. A omissão ou a inclusão de *outliers* depende da área de aplicação específica e da natureza dos dados, incluindo o número e a densidade dos *outliers*.

A interpolação é uma alternativa que pode ser utilizada enquanto os dados não forem muito esparsos. Uma vez que a interpolação assume que existe uma função que une pontos aparentemente adjacentes, se um destes pontos for um *outlier*, a utilização desta proposta pode ser arriscada.

Em nosso estudo, utilizamos uma estratégia de tratamento de dados faltantes híbrida: Devido à existência de 40% de dados faltantes do atributo 'Despesas de Amortização', decidiu-se excluí-lo, pois certamente comprometeria o número de observações disponíveis para o nosso estudo e seus resultados, e a importância atribuída, segundo especialistas, à este atributo para o problema de predição de graus de empresas, a priori, é baixa. O atributo 'Receitas Financeiras' também apresentou dados faltantes para 66 empresas, porém, para estes casos, consideramos o seu valor como zero devido à importância a priori atribuída a este

PUC-Rio - Certificação Digital Nº 0321206/CA

atributo por especialistas. Dessa maneira, ao considerar que a receita financeira de uma empresa é zero, assumimos que estas empresas não possuem rendimentos provenientes de aplicações financeiras.

Mantendo-se os dezoito primeiros atributos (vide tabela 1), e excluindo as empresas que apresentavam dados faltantes em algum(s) deste(s) atributo(s), utilizamos nas seções posteriores desta dissertação um banco de dados completamente preenchido constituído por 318 empresas, descritas no Apêndice 1 e 18 atributos, descritos na tabela 1. Para facilitar a identificação dos mesmos, utilizam-se os códigos apresentados nesta para referenciá-los posteriormente.

Código	Nome do atributo
1	Setor
2	Vendas
3	Receitas
4	EBIT
5	Depreciação
6	Receitas Financeiras
7	Despesas Financeiras
8	Lucro Líquido (conforme) Publicado
9	Lucro Líquido
10	Ativo Total
11	Ativo Corrente
12	Passivo Corrente
13	Passivo Total
14	Endividamento Total
15	Disponibilidade
16	Endividamento Líquido
17	Liquidez Corrente
18	Valor de Mercado

Tabela 1 – Os dezoito atributos utilizados e seus códigos.

A seguir estão descritos o significado destes dezoito atributos:

PUC-Rio - Certificação Digital Nº 0321206/CA

1. **Setor:**

Área de atuação da empresa, a saber: Bens Industriais, Consumo Básico (*Consumer Staples*), Consumo Cíclico (*Consumer Discretionary*), Fontes de Energia, Insumos Básicos, Saúde & Higiene Pessoal, Tecnologia da Informação, Telecomunicações, e Utilidades Públicas;

2. **Vendas:**

Soma que representa o resultado financeiro das operações da empresa, incluindo produtos e/ou serviços em um dado intervalo de tempo, também é conhecido como faturamento total.

3. **Receita:**

Receitas totais da empresa, descontados impostos;

4. **EBIT (Lucro antes de Impostos e Juros):**

Inclui o Lucro total, operacional e não-operacional, antes da dedução de juros e impostos. EBIT é uma métrica tradicional que não inclui custos de capital. Uma de suas características é que é fácil de se calcular sob divisões e sub-divisões de uma empresa;

5. **Depreciação:**

Métrica que exprime a diminuição ou perda de valor dos bens da empresa, tais como equipamentos de escritório e veículos. Terrenos não são considerados como gastos, logo não são considerados depreciativos;

6. **Receitas Financeiras:**

Soma das receitas obtidas através de investimentos financeiros e/ou bancários realizados pela empresa, tais como rendimentos de fundos de investimento, bolsa de valores ou renda fixa. Estas receitas flutuam a cada ano de acordo com a taxa de retorno sobre as aplicações financeiras e o montante investido.

PUC-Rio - Certificação Digital Nº 0321206/CA

7. Despesas Financeiras:

Soma da despesa de juros referentes a todas as obrigações financeiras de uma empresa, sejam elas de curto ou longo prazo, tais como debêntures e empréstimos;

8. Lucro Líquido (conforme) Publicado:

Lucro líquido incluindo itens extraordinários de acordo com convenções de contabilidade dos E.U.A.;

9. Lucro Líquido:

Lucro da empresa, calculado através de faturamento ajustado por custos do negócio, depreciação, juros, impostos, dentre outros;

10. Ativo Total:

Soma do valor de todos dos bens de propriedade da empresa. Inclui fábricas, maquinário, equipamentos e títulos;

11. Ativo Corrente:

Soma dos bens da empresa passíveis de serem convertidos em espécie em 12 meses. Inclui espécie, recebíveis, ações e inventário;

12. Passivo Corrente:

Obrigações financeiras que devem ser pagas em 12 meses, incluindo faturas e compras de bens ainda não pagos, empréstimos de curto prazo e juros resultantes de empréstimos de longo prazo;

13. Passivo Total:

Obrigações financeiras correntes somadas dívidas de longo prazo e impostos resultantes;

PUC-Rio - Certificação Digital Nº 0321206/CA

14. Endividamento Total:

Soma total das dívidas da empresa, incluindo empréstimos de longo prazo e despesas de curto prazo. O Endividamento de uma empresa pode ser benéfico para ela como sendo resultante de compra de fábricas e equipamentos que poderão aumentar sua lucratividade. Porém, um endividamento muito alto pode causar uma despesa financeira muito alta, independentemente de seu faturamento.

15. Disponibilidade:

Total dos fundos em caixa da empresa somados à bens de alta liquidez que podem ser convertidos prontamente em dinheiro, tais como fundos de acionistas, aplicações em overnight e títulos do tesouro;

16. Endividamento Líquido:

Calculado como sendo o endividamento de curto e longo prazos subtraído o caixa. Devido ao caixa ser aplicado contra o endividamento, essa métrica mostra um panorama do endividamento da empresa;

17. Liquidez Corrente:

Razão entre o ativo corrente e despesas de curto prazo, geralmente associada à estabilidade financeira da empresa;

18. Valor de Mercado:

Número total de ações da empresa multiplicado pelo preço corrente de mercado por ação. Esta métrica pode expressar intrínsecamente o tamanho de uma companhia, considerando-se o valor total de todas as ações da empresa. Segundo o Investment Company Institute[2], uma associação norte-americana de investimentos, uma empresa pode ser classificada em grande, média ou pequena de acordo com seu valor de mercado (vide tabela 2).

[2] www.ici.org

PUC-Rio - Certificação Digital Nº 0321206/CA

Valor de Mercado	Bilhões de US$
Grande	5 ou mais
Média	1 a 5
Pequena	Menos que 1

Tabela 2 – Classificação de empresas quanto ao valor de mercado.

2.2
Formatação de dados[3]

Uma vez que esta dissertação se concentrou na classificação e predição de graus de empresa em duas classes (grau de investimento e grau de especulação), agrupou e formatou-se os dados destas empresas nestas duas classes. Destas 318 empresas, 262 são classificadas como grau de investimento e 56 são classificadas como grau de especulação (vide tabela 3).

PUC-Rio - Certificação Digital Nº 0321206/CA

[3] O termo 'formatação de dados' é derivado do conceito mais amplo em inglês *Data Cleansing* que significa além da formatação, a garantia da corretude dos dados, neste caso ratificada pela Thomson Financial Services/BNDES.

Grau	*Ratings*	Quantidade de empresas	Percentual sobre o Total(%)
Investimento	AAA	6	1.89
	AA+	1	0.31
	AA	7	2.20
	AA-	6	1.87
	A+	22	6.92
	A	48	15.09
	A-	38	11.95
	BBB+	41	12.89
	BBB	65	20.44
	BBB-	28	8.81
Especulação	BB+	24	7.55
	BB	11	3.46
	BB-	8	2.52
	B+	5	1.57
	B	5	1.57
	B-	2	0.63
	C	1	0.31
	Total	**318**	**100.00**

Tabela 3 – Distribuição das 318 empresas do BD resultante nas 17 classes de *ratings* e 2 classes de graus.

Os atributos das observações das classes de grau de investimento e grau de especulação mostram-se, em sua maioria, caracterizadas por médias substancialmente distintas (ver tabela 4), porém devido aos altos valores de seus respectivos desvios-padrão, pouco se pode concluir a respeito do seu grau (desfecho) meramente à partir destas métricas;

PUC-Rio - Certificação Digital Nº 0321206/CA

Estatística Descritiva das Empresas					
Código	**Nome do atributo**	**Média (bilhões de US$)**		**Desvio-padrão (bilhões de US$)**	
		Grau de Especulação	**Grau de Investimento**	**Grau de Especulação**	**Grau de Investimento**
2	Vendas	15.19	17.26	33.77	30.26
3	Receitas	1.13	2.18	2.53	3.95
4	EBIT	0.72	2.07	2.84	4.14
5	Depreciação	1.08	0.83	2.67	1.51
6	Receitas Financeiras	0.05	0.04	0.14	0.06
7	Despesas Financeiras	0.68	0.30	1.84	0.79
8	Lucro Líquido (conforme) Publicado	0.05	1.20	1.48	2.85
9	Lucro Líquido	0.06	1.22	1.47	2.77
10	Ativo Total	24.49	21.71	70.99	51.95
11	Ativo Corrente	7.86	6.06	22.87	12.74
12	Passivo Corrente	7.75	5.04	28.16	13.83
13	Passivo Total	20.99	13.51	67.19	40.60
14	Endividamento Total	12.80	5.98	45.38	23.36
15	Disponibilidade	2.81	2.15	8.74	9.67
16	Endividamento Líquido	10.00	3.82	36.83	14.54
17	Liquidez Corrente[4]	1.15	1.00	1.01	0.66
18	Valor de Mercado	7.71	25.49	6.96	43.77

Tabela 4 – Média e Desvios-padrão dos 18 atributos por grau.

[4] Por se tratar de uma razão, à Liquidez corrente não se aplica a unidade descrita na tabela.

PUC-Rio - Certificação Digital Nº 0321206/CA