UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA

KAINAN CREMM RAMOS

MODELAGEM MATEMÁTICA DE RISCO DE CRÉDITO

Campinas

2020

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA

KAINAN CREMM RAMOS

MODELAGEM MATEMÁTICA DE RISCO DE CRÉDITO

Monografia apresentada ao Instituto de Matemática, Estatística e Computação Científica como parte dos requisitos para a obtenção de créditos na disciplina de Projeto Supervisionado II, sob a orientação do Prof. Dr. Laércio Luís Vendite

Campinas

2020

RESUMO

Com a crescente demanda no mercado de crédito, torna-se cada vez mais importante a gestão e o gerenciamento do risco de crédito por parte dos bancos e instituições financeiras. O presente projeto consiste do estudo e da construção de um modelo de Credit Scoring, para mensurar o risco de inadimplência de crédito através do modelo de regressão logística múltipla, visando mitigar os riscos de uma instituição financeiro no que se diz respeito ao financiamento e empréstimo a pessoa física. O modelo construído mostrou-se acurado, com 67,4% de acurácia geral com AUC de 64,7%.

Palavras-chave: Risco de Crédito, Credit Scoring, Regressão Logística.

1 Introdução

Desde a implantação do Plano Real, em 1994, plano que mudou o cenário da inflação descontrolada no Brasil, com a redução expressiva na inflação (que em julho de 1994 beirava 4922% [3] no valor acumulado em doze meses) e a estabilização da moeda, o mercado de crédito se fortaleceu no país.

Após o Real, os bancos e instituições financeiras passaram a expandir ativamente o seu portifólio para a concessão crédito e empréstimos financeiros, visto que o lucro provindo da desvalorização da moeda cessou (ROSA. 2000).

Segundo o Panorama de Crédito do BC (Banco Central), em novembro de 2020, em meio à pandemia do COVID-19, vemos que o saldo de crédito em relação ao PIB brasileiro chega a 53,1%, vs 46,4% no mesmo período em 2019, o que ainda representa uma tendência de crescimento pela demanda de crédito.

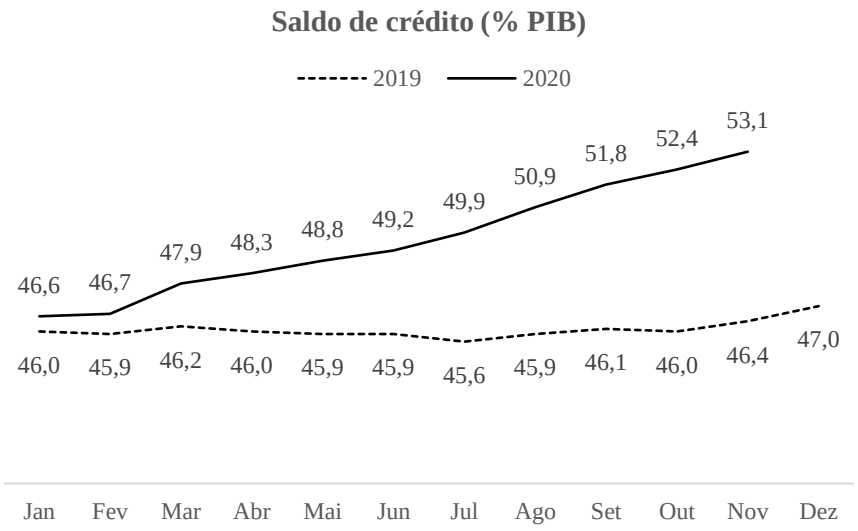

Figura 1: Panorama de saldo de crédito (% PIB) no Brasil

E juntamente com o crescimento da demanda, surgiu-se necessário o controle e gerenciamento do risco de crédito por parte das instituições, buscando mitigar a inadimplência do cliente.

Observando o macro cenário do país, segundo a Peic (Pesquisa de Endividamento e Inadimplência do Consumidor) desenvolvida pela Confederação Nacional do Comércio, temos que 66,3% dos brasileiros estão endividados e 11,2% não terão condições de quitar as suas contas, sendo que os principais tipos de dívida estão voltados para os serviços de cartão de crédito, carnês e financiamento. Reiterando assim a importância dos modelos

de análise de crédito (*Credit Score*), modelos que surgem para classificar os clientes com uma probabilidade de pagar o empréstimo concedido.

Síntese dos Resultados			
	Total de Endividados	**Dívidas ou Contas em Atraso**	**Não Terão Condições de Pagar**
dez/19	65,6%	24,5%	10,0%
nov/20	66,0%	25,7%	11,5%
dez/20	***66,3%***	***25,2%***	***11,2%***

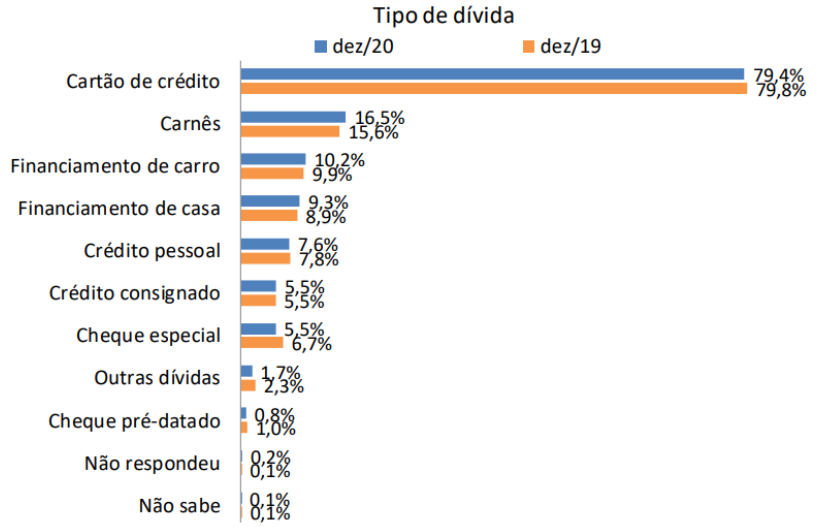

Figura 2: Inadimplência do consumidor e principais dívidas no Brasil – Dez 2020

Com o desenvolvimento e o aprimoramento dos modelos de Credit Score, o que antes era feito de maneira manual por gerentes/analistas nos bancos, de forma altamente suscetível a erros, passou a ser automatizado e com uma inteligência de dados, trazendo maior embasamento, assertividade e agilidade nas decisões.

2 Objetivos

O objetivo de estudo deste é o de:

- Desenvolver um modelo de Credit Score utilizando o Modelo de Regressão Logística Múltipla
- Aplicar o modelo com uma base de dados
- Analisar o poder preditivo do modelo na concessão de crédito

3 Revisão teórica

3.1 Risco de crédito

DOUAT define risco de crédito por:

> Risco de crédito significa o risco de perda em empréstimos ou em investimentos das mais variadas formas que realizamos, devido à falência da empresa investida ou pela deterioração da condição financeira do tomador. Sinteticamente poderíamos afirmar que o risco de crédito está relacionado com a falha em pagamentos durante a vida de uma transação financeira. (DOUAT, 1994; p. 8)

Nesse trabalho, a abordagem do risco de crédito está associada ao risco que uma instituição financeira tem ao conceder crédito a uma pessoa física.

3.2 Regressão logística múltipla

Diante da infinidade de modelos matemáticos e estatísticos para predição e medição de riscos, a regressão logística múltipla é uma das técnicas utilizadas para tratar de variáveis dependentes que são categóricas, ou seja, dicotômicas, binárias. Permitindo a estimativa da probabilidade de ocorrência de um evento, dado um conjunto de variáveis explanatórias (independentes).

Em termos formais:

Seja Y, variável dependente, tal que:

$$Y_i = \begin{cases} 1, \text{ se o evento E ocorre} \\ 0, \text{ caso contrário} \end{cases}$$

E o vetor $X_i = \{x_{i1}, x_{i2}, \ldots, x_{ij}\}$, das j variáveis independentes (discretas ou contínuas), o modelo de regressão logística é descrito como:

$$P\{Y_i = 1\} = f\left(\beta_0 + \sum_j \beta_j x_{ij} + \varepsilon\right)$$

Onde:

- β_i são os coeficientes do modelo, calculados via máxima verossimilhança

- x_{ij} representam o estado das variáveis independentes, associadas a i-ésima variável dependente Y
- ε é o erro do modelo
- f é função logística, dada por $\mathrm{f}(x) = \frac{e^x}{1+e^x}$,

Podemos ver o comportamento da função na figura 2, cujo comportamento probabilístico tem o formato da letra S, tendo aplicações em diversas áreas: na medicina, em modelos de crescimento de tumores; na biologia, em dinâmicas de população, etc.

Na forma linearizada, o modelo é descrito pela função logit, sob a qual aplicamos os métodos clássicos para determinação dos parâmetros β_i.

$$logit(P\{Y_i = 1\}) = \beta_0 + \sum_j \beta_j x_{ij} + \varepsilon$$

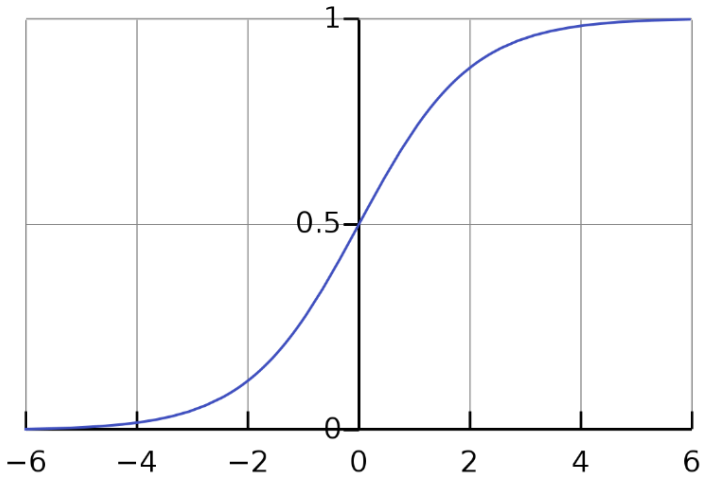

Figura 3: O comportamento da função logística

Ao construir o modelo, por Y se tratar de uma variável categórica, define-se um critério de classificação δ, a partir do qual consideramos a ocorrência do evento (Y=1) ou não; de forma que a mudança nesse parâmetro muda a eficácia do modelo.

$$Y = \begin{cases} 1, \text{ se } P(Y=1) \geq \delta \\ 0, \text{ caso contrário} \end{cases}$$

4 Credit Scoring com a Regressão Logística Múltipla

A partir do modelo teórico, definimos a variável dependente Yi a ser prevista, referente ao cliente i, como:

$$Yi = \begin{cases} 1, \text{ quando o cliente é inadimplente (apresentou inadimplência de } 90+ \text{ dias)} \\ 0, \text{ caso contrário} \end{cases}$$

Sendo a matriz dependente, referente a base de dados, correspondente a x_{ij} a informação j do cliente i:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}$$

4.2 Descrição das variáveis

A base de dados utilizada é pública, disponibilidade no repositório da Universidade da Califórnia [1].

Com um total de m clientes ($m = 1000$), de onde temos 300 deles que são inadimplentes (Y=1, isto é, tiveram inadimplência de 90 ou mais dias), sendo os restantes 700 adimplentes (Y=0).

Os n campos ($n = 10$) de cada cliente são qualitativos e numéricos, contidos na tabela 1.

Nome	Categoria
Risk (Y)	Alto risco de inadimplência (1), Baixo risco (0)
ID	Identificador do cliente (String)
Age	Idade (Inteiro)
Sex	Gênero (“Masculino” / “Feminino”)
Job	Emprego (1 – não qualificado, 2 – qualificado, 3 – muito qualificado)
Housing	Tipo de habitação (“casa própria” / “aluguel”)
Saving accounts	Poupança (Deutsch Mark)
Checking account	Conta corrente (“little” / “moderate” / ”rich”)
Credit amount	Quantidade de crédito (DM)
Duration	Duração do empréstimo (meses)
Purpose	Razão do crédito (carro, educação, reparos, viagem, etc)

Tabela 1: Variáveis de interesse do problema

4.3 Tratamento das variáveis

Com a *database*, o tratamento das variáveis consistiu de:

- Limpeza da base: ajuste dos termos não-nulos.
- Transformação das variáveis categóricas explicativas através de uma função *Encoder*. Gerando assim um valor numérico para a interpretação do modelo.
- Normalização das variáveis, garantindo que a escala não comprometa a ponderação dos pesos do modelo. Sobretudo pois campos como Quantidade de crédito (Credit amount), apresentam valores em marco alemão (antiga moeda da República Federal da Alemanha) que são bem distintos em escalas dos demais, que em sua maioria são variáveis categóricas, como o valor em Conta corrente (Checking account).

4.4 Análise exploratória

Traçando-se as métricas estatísticas e distribuições de probabilidade, analisamos o perfil da população, visando compreender as variáveis a serem efetivamente incluídas na regressão, visando ter um maior poder explicativo, minimizando o erro do modelo.

A figura 4 ilustra as distribuições plotadas, sendo que em laranja temos a distribuição para os clientes inadimplentes e a azul para os adimplentes. De modo que as diferenças entre as distribuições (não somente em termos das métricas, mas também do comportamento das curvas), a priori, nos mostram relações de significância.

Em linhas gerais, podemos observar:

- Quanto à duração do empréstimo/financiamento de crédito (Duration), podemos ver o risco de inadimplência é maior para dívidas a um maior prazo, o que é intuitivo visto que as taxas de juros para o contratante são proporcionalmente maiores, podendo acarretar no endividamento e consequentemente no não-pagamento da transação.
- Observando o fator de gênero, temos ambas as curvas bem próximas e com formatos semelhantes, o que nos dá indícios de que essa variável possa não ter relevância para o modelo.

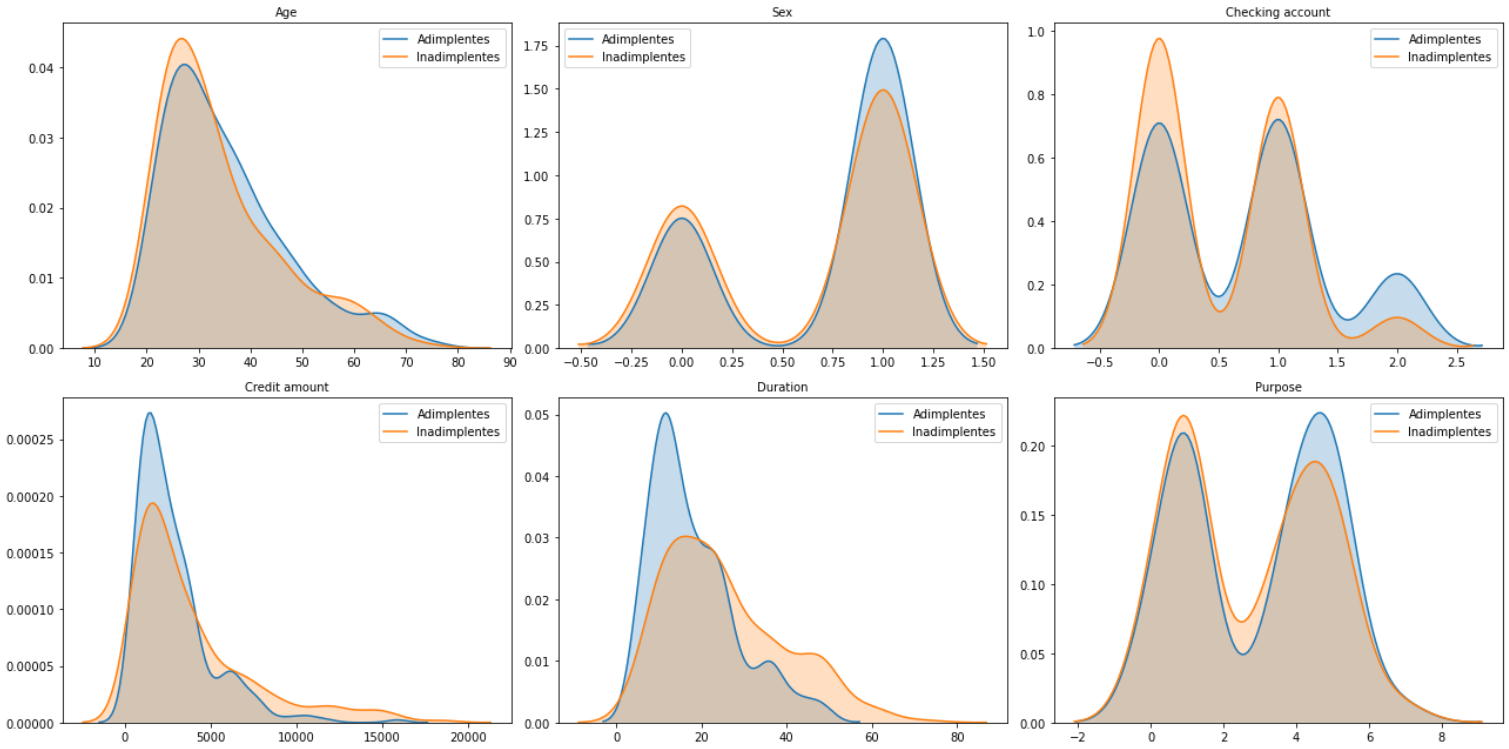

Figura 4: Distribuições de probabilidade das variáveis de interesse

4.5 Aplicação do Modelo

Com o pré-processamento e o balanceamento da base, de modo a ter uma boa proporção entre as variáveis categóricas da amostra (sem com que o modelo tendesse a dar muitos "alarmes falsos"), separamos os grupos de treino e teste utilizando as proporções de 50%-50%.

Finalmente, com o ajuste dos coeficientes β_i do modelo, validamos a predição no conjunto teste, cujos resultados encontram-se na seção seguinte.

5 Resultados

A tabela 2 apresenta as métricas para o modelo e a figura 5 a Matriz de Confusão com os verdadeiros e faltos positivos/negativos.

	Precision	Recall	F1 - score
1	0.722	0.678	0.699
0	0.621	0.669	0.644

Tabela 2: Métricas do modelo

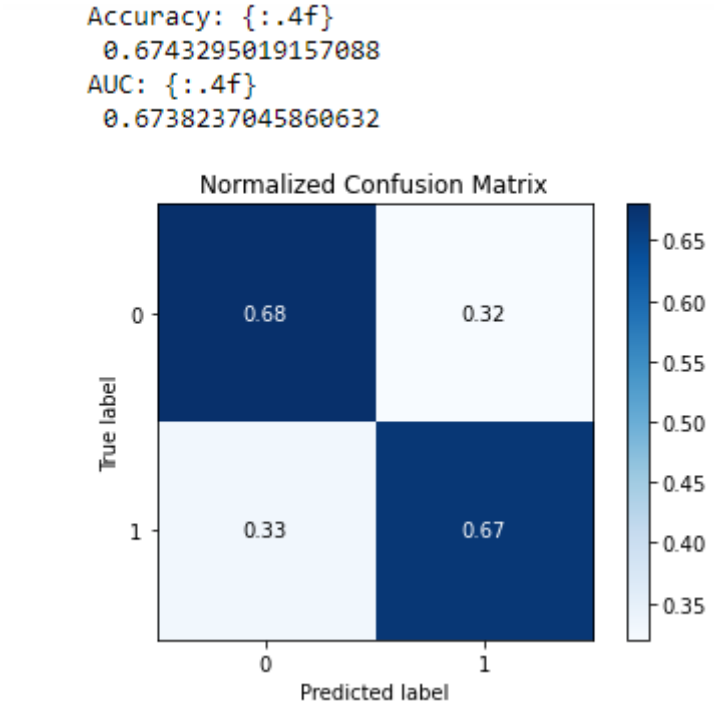

Figura 5: Matriz de confusão do Modelo construído

- AUC: 67,3%

Como medida de separabilidade, o AUC (valor entre [0,1]: 1 caso ideal, 0 pior caso) nos mostra o desempenho do modelo utilizando como base a taxa de falso positivo e a taxa de verdadeiro positivo. Dado o threshold de 50%, temos um valor coerente.

- Acurácia: 67,4%

A acurácia nos mostra, em média, a frequência da assertividade do nosso modelo; sendo que o percentual atingido se mostra pertinente. Além de que, observando a figura 5, vemos que o modelo apresentou 67% de taxa de verdadeiros positivos e 68% de verdadeiros negativos, um comportamento esperado nas diagonais principais.

Em literatura consultada, vemos que existem diversos outros métodos utilizados para Credit Scoring, tais como Árvores de Classificação, PL, Redes Neurais e Algoritmo

Genéticos, sobretudo também vemos como a acurácia média varia bastante, conforme indica a tabela 3 [2].

	Regressão Linear	Regressão Logística	Árvores de Classificação	Programação Linear	Redes Neurais	Algoritmos Genéticos
Henley(1995)	56,6	**56,7**	56,2	-	-	-
Boyle (1992)	**77,5**	-	75	74,7	-	-
Srinivisan(1987)	87,5	89,3	**93,2**	86,1		
Yobas (1997)	**68,4**	-	62,3	-	62	64,5
Desai(1997)	66,5	**67,3**	**67,3**	-	64	-

Tabela 3: Precisão dos modelos de Credit Scoring para demais referências

E por mais que cada caso tenha sua especificidade da base de dados, técnicas heurísticas e métodos empregados, quando olhamos para a Regressão Logística, vemos que a variação da acurácia vai de 56,7% a 89,3%, sendo o nosso valor muito próximo do que foi desenvolvido estudo de Desail (1997).

6 Conclusão

O projeto consistiu do estudo e da construção de um modelo de Credit Scoring, para mensurar o risco de inadimplência de crédito através do modelo de regressão logística múltipla, visando mitigar os riscos de uma instituição financeiro no que se diz respeito ao financiamento e empréstimo a pessoa física. O modelo construído mostrou-se acurado, com 67,4% de acurácia geral com AUC de 64,7%.

Por fim, reitera-se a importância do desenvolvimento e estudo de modelos cada vez mais robustos de Credit Scoring, especialmente no Brasil, onde o hub de fintechs (startups do setor financeiro/crédito) tem crescido exponencialmente nos últimos anos, chegando a triplicar empréstimos na Pandemia do COVID-19[1]. De forma a contribuir tanto para as instituições, como também para a estabilidade e saúde econômica do país.

[1] Fonte: ABCD – Associação Brasileira de Crédito Digital

7 Referências

[1] **German Credit Data Dataset**. Machine Learning Repository. Disponível em: *https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)*

[2] Gonçalves, Eric. **Análise de Risco de Crédito com uso de Modelos de Regressão Logística, Redes Neurais e Algoritmos Genéticos**. Dissertação de Mestrado, FEA USP, 2015.

[3] **Plano Real**. Banco Central do Brasil, 2018. Disponível em: *https://www.bcb.gov.br/controleinflacao/planoreal*

[4] **Regressão Logística**. E-Disciplinas USP. Disponível em: *https://edisciplinas.usp.br/pluginfile.php/3769787/mod_resource/content/1/09_RegressaoLogistica.pdf*

[5**] Panorama Crédito**. Federação Brasileira de Bancos, 2020. Disponível em: *https://cmsportal.febraban.org.br/Arquivos/documentos/PDF/Panorama%20Boletim%20de%20Cr%C3%A9dito_fev%2020.pdf*

[6] Campos, Cristina Ana. **Número de brasileiros com dívidas cresce no fim de 2020**, Agência Brasil, 2020. Disponível em: *https://agenciabrasil.ebc.com.br/economia/noticia/2021-01/numero-de-brasileiros-com-dividas-cresce-no-fim-de-2020*

[7] ROSA, P.T.M. **Modelos de credit scoring: Regressão Logística**, CHAID e Real. Dissertação de Mestrado, Departamento de Estatística,Universidade de São Paulo, São Paulo, 2000. 15

[8] Pesquisa de Endividamento e Inadimplência do Consumidor (Peic)**. Confederação Nacional do Comércio de Bens, Serviços e Turismo, 2020.**

[9] **Pandas documentation**. Pandas, 2020. Disponível em: *https://pandas.pydata.org/pandas-docs/stable/index.html*

[10] Moura, Machado Gabriela. **Regressão logística aplicada a risco de crédito**. Monografia, Departamento de Matemática Aplicada, FURG, 2018.

[11] DOUAT, J. **Desenvolvimento de Modelo para Administração de Carteira de Crédito**. São Paulo, 1994. Tese de Doutorado – FEA USP, 1994.