5CONTECSI June 04 - 06, 2008 São Paulo, Brazil

5th International Conference on Information Systems and Technology Management
5º Congresso Internacional de Gestão da Tecnologia e Sistema de Informação

De 04 a 06 de Junho de 2008 - São Paulo - Brasil

PS-1079

DATA MINING IN BUSINESS MANAGEMENT: AN APPLICATION IN SALES OF TELECOMMUNICATION SERVICES IN THE CORPORATE SEGMENT

Teófilo Camara Mattozo, TELEMAR NORTE LESTE S/A - RN - Brasil mattozo@oi.com.br
José Alfredo Ferreira Costa, Laboratório de Sistemas Adaptativos, Centro de Tecnologia, Universidade Federal do Rio Grande do Norte - RN - Brasil alfredo@dee.ufrn.br
Manoel Veras de Sousa Neto, Departamento de Administração, Universidade Federal do Rio Grande do Norte - RN - Brasil manoel.veras@uol.com.br

Telecommunications are one of the most dynamics and strategic areas in the world. In a complex and competitive scenario with pressure for better results with fewer resources, there is a constant need in the organizations for searching new and improved forms of administration with better understanding of the processes and customers. Techniques such as knowledge discovery in databases (KDD) and data mining (DM) appear as alternatives for enabling transform raw and complex data in models that may allow better decision making. This paper presents an application of multivariate regression analysis for modeling and explaining sales in the segment of corporate telecommunications, through acting indicators. A series of comparative analyses were accomplished with intend to identify the most appropriate or impacting indicators to perform the best decisions when selling corporate services. It is also presented a systemization of KDD activities which integrates the methodologies such as CRISP-DM and SEMMA in an interactive design framework. Variable selection methods and statistical validation of the model are presented. Results are presented from real databases of large Brazilian telecommunications company.

Keywords: Performance Indicators; Knowledge Discovery in Database; Business Management; Telecommunications; Decision Support Systems.

APLICAÇÃO DE MINERAÇÃO DE DADOS NA GESTÃO DE VENDAS DE SERVIÇOS CORPORATIVOS DE TELECOMUNICAÇÕES

Telecomunicações é uma das mais dinâmicas e estratégicas áreas no mundo atual. Em um cenário complexo e competitivo com pressão para resultados melhores com menos recursos, há uma necessidade constante nas organizações de procurarem novas formas de gerenciamento com melhor compreensão dos processos e dos clientes. A existência de bases de dados nas empresas passou a ter maior importância. Técnicas como a descoberta do conhecimento em bases de dados (DCBD) e mineração de dados (MD) aparecem como alternativas para permitir a transformação de dados brutos e complexos em modelos que podem ajudar nas tomadas de decisões. Esse artigo apresenta uma aplicação de análise de regressão multivariada para explicar vendas no segmento de telecomunicações corporativo, a partir de uma série de indicadores de desempenho. Análises comparativas foram realizadas com objetivo de identificar os indicadores que possuem maior impacto no modelo e que permitam auxiliar na tomada de melhores decisões no momento da venda no segmento corporativo. É apresentada também uma sistematização de atividades de DCBD que integra metodologias tais como CRISP-DM e SEMMA em um ambiente interativo. Métodos de seleção de variáveis e a validação estatística do modelo são apresentados. Os resultados apresentados são baseados em bases de dados reais obtidos em uma grande companhia brasileira no setor de telecomunicações.

Palavras-chave: Indicadores de Desempenho; Sistemas de Apoio a Decisão; Descoberta de Conhecimento em Bases de Dados; Gestão de Negócios em Telecomunicações.

1 – INTRODUÇÃO

As empresas industriais vêm enfrentando recorrentemente situações adversas oriundas do aumento da competitividade, da estagnação de mercados, da dependência de uma economia instável e globalizada, e principalmente de clientes mais exigentes e sofisticados. Pressões do lado dos custos (aumento da competitividade) e dos preços (clientes mais exigentes) conduzem a empresa ao gerenciamento da rentabilidade, e não mais apenas da receita e da participação de mercado. Dentro desse contexto, o mercado industrial (também conhecido como mercado empresarial, *business to business* ou ainda mercado organizacional) tem sido fortemente marcado pela necessidade de se buscar e aplicar novas técnicas e ferramentas de gestão. Essas técnicas e ferramentas têm o objetivo de traduzir, em linguagem organizacional corrente, o conceito de excelência empresarial, em uma perspectiva prática, que proporcione soluções aos desafios organizacionais. Um dos desafios que as organizações necessitam superar é o de descobrir como aperfeiçoar a sua produtividade das linhas de negócios, otimizando a execução dos principais processos.

O estabelecimento de relacionamentos entre os parâmetros de desempenho dos processos é um grande desafio nas empresas. O mapeamento dessas relações objetiva identificar as medidas de desempenho direcionadoras ao resultado e de aumentar a capacidade preditiva das medidas de desempenho (KENNERLEY & NEELY, 2003). Diversos modelos de medição de desempenho surgiram a partir da década de 80, resultantes das mudanças gerenciais

ocorridas, entre eles: *SMART - Performance Pyramid (*CROSS & LYNCH*,* 1990*); Balanced Scorecard (*KAPLAN & NORTON*,* 1997*); Integrated Performance Measurement System* (BITITCI et al., 1997)*; Integrated Dynamic Performance Measurement System* (GHALAYINI, et al., 1996) e *Performance Prism* (GHALAYINI, et al., 1996).

O acompanhamento do desempenho do processo de planejamento, por exemplo, deve ser realizado a partir da definição de indicadores de desempenho, que facilitem a análise das causas e efeitos dos desvios entre o planejado e o realizado, de forma que os gestores possam corrigir distorções na execução do plano. Nesse contexto, passa a ser muito importante a utilização de indicadores que realmente possam verificar se a missão da empresa está sendo atingida. Para isso, a amplitude dos sistemas de controle de desempenho tem sido modificada a fim de incluir os ativos intangíveis em seu conjunto de indicadores, além dos resultados financeiros. Diante disso, o problema é saber se as medidas de desempenho escolhidas são as mais adequadas, se o objetivo está sendo alcançado e saber se as melhorias implantadas estão surtindo efeito. Normalmente, não há informações suficientes ou necessárias para responder a essas questões. Desse modo, a correta definição dos indicadores de desempenho se torna ponto crucial para o sucesso de uma empresa, já que esses podem ser usados para acompanhar os resultados obtidos alinhados às estratégias em diferentes níveis da organização.

Por outro lado, constata-se um crescimento substancial da quantidade de dados armazenados pelas organizações. A análise desses dados, produzidos e armazenados em larga escala, realizada por especialistas através de métodos manuais tradicionais, torna-se impraticável. A informação obtida e a sua utilização vêm desempenhando um papel fundamental na geração de valor dessas organizações. Mas a transformação dessas grandes quantidades de dados em informação não é uma tarefa trivial. Informações relevantes não são facilmente obtidas a partir de sistemas de banco de dados, uma vez que, boa parte dos dados operacionais provenientes desses sistemas, não apresentam relevância quando estudados individualmente (O'GUIN et al., 2001). Surge então à necessidade de explorar melhor os dados gerados de forma individual por esses grandes sistemas para extração de conhecimento implícito e utilizá-los no âmbito da solução de problemas empresariais.

A extração de conhecimento, conhecimento descrito aqui como informação dotada de um certo contexto, é uma área em constante desenvolvimento, quer no que se refere ao refino das técnicas e ferramentas de Mineração de Dados (MD), quer na melhoria das tecnologias complementares que contribuem para a persecução de projetos. As áreas de aplicação são inúmeras e em vários casos existem ganhos significativos em problemas de decisão, cada vez mais sofisticados. Apesar dos sucessivos progressos tecnológicos e do aprofundamento das competências dos especialistas, a concretização de projetos de descoberta de conhecimento continua a ser difícil e a exigir um conjunto bastante diverso de conhecimentos, ações e decisões (ARNETT et al., 2000). Descobrir conhecimento significa extrair, de grandes volumes de dados, informações relevantes e até então desconhecidas, que se revelam úteis e válidas para processos de tomada de decisão. Recorrendo à definição elaborada por Fayyad et al. (1996), a descoberta de conhecimento em bases de dados (DCBD) pode ser definida como "um processo interativo não trivial de identificar novos padrões nos dados que sejam válidos, potencialmente úteis e interpretáveis".

Um dos muitos desafios que emerge nesse processo consiste na seleção dos métodos de MD mais apropriados para determinada aplicação. Não existem critérios simples e gerais que suportem sistematicamente tal decisão. A adequação das técnicas de MD varia em função de diferentes tipos de fatores, entre os quais alguns complexos (ex. pressupostos dos métodos em termos de características dos dados), subjetivos (ex. simplicidade na interpretação de resultados) e às vezes conflituosos (ex. nível de precisão versus facilidade de interpretação de resultados) (ZHOU, 2003).

Esse artigo pretende identificar relações entre parâmetros de venda e indicadores de desempenho, visando descobrir eventuais novos conhecimentos, por intermédio de técnicas de DCBD, a partir de dados históricos da organização em estudo. Focaliza-se a área de administração de vendas, área que normalmente produz uma série de dados relativos ao relacionamento com diversos clientes e que na maioria das vezes não são devidamente tratados. Objetiva-se nessa pesquisa subsidiar a previsão da receita gerada pelas vendas de telefonia fixa, melhorando o seu grau de confiança, com a intenção de auxiliar os gestores na tomada de decisões. Utilizou-se análise de regressão multivariada para esse caso na tentativa de explicar o comportamento das vendas no segmento de telecomunicações corporativo, a partir dos indicadores de desempenho utilizados. É apresentada também uma sistematização de atividades de DCBD que integra as metodologias CRISP-DM e SEMMA em um ambiente interativo. Métodos de seleção de variáveis e a validação estatística do modelo são apresentados. Os resultados apresentados são baseados em bases de dados reais obtidos em uma grande companhia brasileira no setor de telecomunicações.

2 - METODOLOGIA IMPLEMENTADA

A estratégia de desenvolvimento desse trabalho obedece a um conjunto de premissas que vão desde a necessidade de fixação dos objetivos para as atividades a serem realizadas (LO, 2002), a determinação das fontes de dados internas e externas à organização (HUGHES, 1995; COOKE, 2000) e a capacidade tecnológica de processar grandes volumes de dados, capazes de suportar atividades de DCBD (ZWICK et al., 2004). A pesquisa visa, conforme dito anteriormente, auxiliar a área de Gestão de Negócios em Telecomunicações a usar o conhecimento extraído das Bases de Dados (BD) nas suas atividades de gestão, as quais se encontram associadas a objetivos de planejamento mais vastos, que por sua vez correspondem ao reflexo da estratégia organizacional (MATTOZO, 2007).

O modelo utilizado integra diferentes abordagens sendo baseado principalmente na exploração dos conceitos e características do CRISP-DM (CHAPMAN et al., 2000), cruzando-o quer com as atividades de vendas, quer com as questões inerentes à integração dos modelos de MD (referido aqui como componente integrante do processo de DCBD).

O caso da pesquisa descrita nesse trabalho desenvolveu-se em uma grande empresa de telecomunicações no Brasil que comercializa produtos e serviços de telefonia básica, avançada, comunicação de dados, longa distância nacional e internacional e mobilidade. Foram coletados dados de 15 empresas, no período de janeiro a dezembro de 2006, com atuação no Nordeste e Sudeste do Brasil as quais são agentes autorizados e comercializam todos os produtos voltados para a área empresarial. Como apoio a execução da pesquisa foi utilizado o *software* de Mineração de Dados SPSS (SPSS, 2007).

O modelo utilizado possui três fases: *estudo do negócio, extração de conhecimento e avaliação dos resultados obtidos*. Na primeira fase considera-se que os dados podem ser obtidos a partir de diferentes fontes. Após o seu registro e análise é então criada uma base de dados de administração de vendas, com vista ao suporte às fases seguinte, correspondente à extração de conhecimento. A aplicação dos resultados obtidos é concretizada na fase de aplicação a atividades de administração de vendas.

Três questões foram fundamentais para o desenvolvimento da pesquisa:

- O levantamento de informações sobre a empresa de forma a criar um modelo, para definir as suas necessidades, os fluxos de informações e dados vitais para o atendimento dos objetivos de negócio (WIRTH, 2000);
- A definição do banco de dados. A sua escolha vai depender do tamanho, da quantidade de variáveis presentes e das áreas organizacionais envolvidas. Um *data warehouse* ou um *data mart* consegue lidar com dados armazenados, complicados e complexos, mas existe também possibilidade, por questão de custo e simplicidade, de aplicar as ferramentas de Mineração de Dados diretamente em bancos de dados relacionais (FEELDERS et al., 2000);
- A escolha das técnicas e ferramentas de Mineração de Dados. A escolha da técnica vai depender dos tipos de dados existentes (por exemplo, dados numéricos e não-numéricos) e também de qual tipo de relacionamento deve ser pesquisado entre as medidas de desempenho. Por exemplo, uma relação de causa-e-efeito requer uma comprovação a partir de cálculos matemáticos. Já a relação lógica é baseada em deduções sobre as relações existentes. Vale observar que ambas usam técnicas de Mineração de Dados diferentes (WELGE, 2001). Dessa forma, é possível estabelecer, a partir de uma grande massa de dados históricos sobre o desempenho, o relacionamento entre as medidas de desempenho de um Sistema de Medição de Desempenho (SMD).

Após revisão bibliográfica realizada foi possível identificar os requisitos a cumprir no desenvolvimento da pesquisa para verificação dos objetivos propostos descritos.

3 - ESTUDO DE NEGÓCIO

Definir um estudo pode envolver articular uma meta, escolher uma variável dependente ou uma saída que caracterize um aspecto da meta e especificar os campos de dados que são usados no estudo. Por outro lado, a meta pode ser usada para agrupar tipos similares de dados ou para identificar exceções em um conjunto de dados (ADRIAANS et al., 1996). A identificação de exceções é geralmente usada na descoberta de fraude ou de dados incorretos.

Para estabelecer os relacionamentos entre as medidas de desempenho usando a Mineração de Dados, é necessário conhecer o sistema de medição de desempenho (SMD) e os sistemas de informação da empresa onde o método proposto vai ser aplicado. O levantamento das informações abrange três aspectos na definição do modelo conceitual (DE TONI & TONCHIA, 2001):

- Mapeamento dos possíveis relacionamentos entre as medidas de desempenho;
- Identificações de quais informações são relevantes para os tomadores de decisão e que eles ainda têm dificuldades em obtê-las;
- Levantamentos de como os sistemas de informação da empresa estão estruturados, principalmente na questão do armazenamento dos dados a fim de identificar quais são as melhores possibilidades de aplicação da MD.

Dentre as várias ferramentas ou sistemas de medição de desempenho, o *Balanced ScoreCard* é um modelo de gestão que auxilia a empresa a traduzir a estratégia em objetivos, indicadores, metas e planos de ação, balanceados e alinhados, que direcionam comportamentos e performances (KAPLAN & NORTON, 1997). Tal como indicadores registram o dia-a-dia dos processos associados, o BSC o faz com a estratégia da empresa. Dentro da perspectiva do BSC a estratégia é tarefa de todos, cada um com sua parcela do desdobramento de metas corporativas. Isso é obtido com o alinhamento dos indicadores ao mapa estratégico e a sua gestão nos diversos níveis hierárquicos da empresa. A gestão por indicadores alinhada ao BSC permite determinar qual a contribuição dada por cada unidade de negócio no resultado global da Empresa (NORREKLIT, 2000).

Os atributos do BSC nas empresas analisadas incluem aprendizado e crescimento, capital informacional, capital organizacional, concorrência, excelência operacional, financeira, foco do cliente, gestão de parceiros, inovação e ação e mercado. Esse trabalho focalizou na Unidade de Negócios Empresarial, os seguintes indicadores de desempenho associados com os seguintes atributos do BSC, quais sejam:

- Concorrência: Vendas x Mapeamento da Concorrência;
- Excelência Operacional: Assertividade de Entrega e Cadastro de Contratos, Retorno de Contratos, Abertura de Ordem de Serviços, Conversão de Contratos e Cancelamento de Vendas;
- Foco do Cliente: Média de Visitas por Consultor, Assertividade de Agendamento e Visitação versus Agendamento;
- Gestão de Parceiros: Produtividade de Visita, Eficácia de Visitas, Mix de Produtos, Distribuição de Vendas e Produtividade de Oportunidades;
- Inovação e Ação: Volume de Vendas Customizadas, Aprovação de Viabilidade e Assertividade de Análise de Viabilidade.

Esses indicadores estão contemplados em todo fluxo de venda, ou seja:

- Pré-visita através da oportunidade de vendas, agendamento de visitas e no planejamento da visita;
- Visita pelo registro da visita através da ficha de visita eletrônica;
- Pós-visita na emissão de propostas e contratos e acompanhamento da abertura da ordem de serviços de instalação.

4 - EXTRAÇÃO DE CONHECIMENTO DE BASES DE DADOS

O desenvolvimento do processo de administração de vendas suportado na DCBD concretiza-se pela realização das atividades de análise e exploração dos dados, pré-processamento, modelagem e avaliação de resultados (UTHURUSAMY & FAYYAD,

2002). O processo de extração de conhecimento ajuda a decodificar as relações existentes entre os dados e que estão para além da capacidade cognitiva do analista (LO, 2002). O objetivo da DCBD no âmbito da sua aplicação em projetos de administração de vendas é transformar dados em resultados práticos permitindo numa fase seguinte atuar com a informação obtida.

A preparação dos dados é uma etapa de grande importância para todo o processo de DCBD que engloba a seleção dos dados, limpeza e transformação. Para o sucesso do processo de DCBD é necessário que os dados tenham sido corretamente selecionados, corrigidos e transformados. São estudadas e aplicadas as estratégias para tratamento de dados incorretos, além de alternativas para tratar os registros com dados faltosos ou incompletos (WIRTH, 2000). O processo de DCBD desenvolve-se segundo o processo da metodologia utilizada, no qual existem as fases de análise de dados (avaliação da qualidade); pré-processamento; modelagem e finalmente avaliação dos modelos obtidos.

A existência de uma BD com grandes dimensões evidencia uma das maiores limitações inerente a sua utilização na maioria das organizações que é a (in)capacidade para extrair informação relevante, para além daquilo que os processos tradicionais permitem. O estudo partiu de uma BD com mais de 200.000 registros permitindo ainda concretizar as práticas enumeradas na administração de vendas de serviços de telecomunicações, com resultados práticos relevantes.

Foi necessário um aprofundamento no sistema de informação da empresa, principalmente o que se relaciona às decisões existentes, de modo a verificar quais são os bancos de dados transacionais e como eles são utilizados pelos tomadores de decisão. Esse entendimento auxiliou na utilização das fontes de dados possíveis para o uso das técnicas e ferramentas de Mineração de Dados. Durante a fase de avaliação dos dados internos disponíveis, foram conhecidos os bancos de dados disponíveis para esse estudo de caso e os dados armazenados. Foram levados em consideração os critérios técnicos delineados na definição dos objetivos, para que o conjunto de dados resultante esteja apropriado ao restante do processo.

Os dados internos disponíveis provinham de dados armazenados em diferentes sistemas de informação da Empresa. O processo de obtenção desses dados envolveu ainda vários recursos em diferentes níveis hierárquicos da empresa como gestores ou administradores de sistemas de informação, no sentido de se garantir o acesso aos mesmos dados. A empresa dispõe de diversos sistemas para registro do seu dia-a-dia operacional, porém tais dados são armazenados em bancos de dados distintos.

Para esse estudo, foram disponibilizados os dados do SA-3, ASE e SCI migrados para o formato do banco de dados Microsoft Access, além de outros documentos com dados relevantes ao processo. Contudo, quando se utilizam bases de dados distintas é necessário que os dados sejam migrados para um formato compatível ou que os algoritmos de extração estejam preparados para trabalhar com bancos de dados heterogêneos. Um resumo da descrição dos sistemas de dados é feita a seguir:

- *Sistema de Vendas (SA 3)*: Permite a equipe de vendas do Empresarial o controle e gerenciamento sobre todo o *workflow* de vendas, possibilitando a gestão das oportunidades, gestão de atividades, planejamento de visitas, mapeamento da concorrência, ações comerciais, medição da carteira, parâmetros de desempenho e, conseqüentemente, análise do funil de vendas, além relatórios gerenciais bem estruturados e detalhados;
- *Sistema de Cadastro de Informações (SCI)*: Esse sistema armazena informações das atividades de todos os serviços de telecomunicações comercializados pelos consultores, abrangendo toda a parte contratual, ou seja: visão de contrato e suas condições, identificação dos serviços comercializados, condições de aceite para liberação de solicitação de abertura de ordem de serviço, parâmetros de desempenho, etc.;
- *ASE Empresarial*: Sistema gerencial que disponibiliza informações relacionadas a ações comerciais, carteira de cliente segmentada, priorização de visitas a serem realizadas, entre outras informações correlatas.

A constituição da BD inicial dos parâmetros de vendas de telecomunicações foi feita com elaboração de uma lista prévia organizada pelo tipo de informação a que cada registro correspondia, informação do cliente, do serviço comercializado (em nível de detalhes), oportunidades identificadas, mapeamento dos serviços existentes disponibilizados pela concorrência (com registro dos serviços, prazo contratual, data de vencimento, etc.) bem como várias outras informações pertinentes ao negócio. Essa lista foi criada a partir da importação do atributo, *Agente Autorizado*, de BD relativas às vendas realizadas.

Após a disponibilidade dos dados internos a partir de fontes diversas, usando como identificador único o atributo *Agente Autorizado* (chave), foi realizado um processo de unificação dos dados devido à elevada redundância de informação possível à mesma comercialização, pois se pode encontrar registrado em mais do que uma BD. Nesse caso após a importação das diferentes tabelas, procedia-se a uma ordenação da tabela resultante pelos atributos em análise e manualmente o pesquisador poderia detectar a ocorrência de registros duplicados. Ocorreram poucas situações com duplicação de registros decorrentes da concentração de tabelas de BD distintas proporcionando a coexistência dos mesmos atributos, com o mesmo significado, mas codificados de modo diferente, provocando redundância de informação.

A BD sobre a qual se desenvolveu todo o trabalho de pesquisa resultou da operacionalidade do sistema de administração de vendas. Os dados guardados na BD foram armazenados em tabelas distintas:

- *Parâmetros de Desempenho:* dados relativos às visitas realizadas, agendadas, informações da concorrência, ações comerciais, entre outras;
- *Transações Comerciais:* dados transacionais, relativos a informações sobre a emissão e cadastro de contratos dos serviços de telecomunicações comercializados.

Após obtenção dos dados uma etapa intermédia, importante e que demandou grande parcela de tempo, foi a seleção dos registros. Face à heterogeneidade das fontes de dados foi necessário proceder a uma uniformização dos mesmos, no sentido de evitar a duplicação de registros, incongruências, inconsistências e violações de domínio.

Embora pese o esforço desenvolvido na filtragem e limpeza de dados, expresso anteriormente, a base apresentou problemas, principalmente em termos relativos ao processo de entrada de dados, dados omissos ou incorretos, os quais tiveram que ser tratados. A ocorrência desse tipo de problema justifica-se basicamente pela seguinte razão fundamental: o processo de entrada dos dados das vendas realizadas é manual, proporcionando a inserção de valores incorretamente ou a falha na interpretação dos dados inscritos manualmente pela área de pós venda.

A existência de valores em branco num determinado atributo suscita o tratamento desse atributo, com duas opções distintas, mas viáveis: eliminação do atributo da BD (no caso de valor omisso da maioria dos registros) ou processamento do atributo, isoladamente ou em função de outros. Nesse trabalho a abordagem adotada nesse último caso foi o preenchimento dos casos omissos com valores equivalentes à média desse atributo em todos os registros de uma determinada condição do serviço. Isso ocorreu poucas vezes na segmentação do serviço comercializado bem como, na entrada dos dados das visitas realizadas. Entretanto, esses fatos não foram relevantes na constituição da BD final dos parâmetros de venda, pois, na derivação de novas variáveis, essas informações não faziam parte da mesma.

Verificou-se com pouca freqüência a ocorrência de valores anormais para alguns atributos, em alguns registros. Ao contrário do processo de tratamento de dados omissos, os atributos encontravam-se preenchidos, mas com a possibilidade de não corresponderem à realidade. Essa fase procurou recuperar a integridade do registro pelo recurso do tratamento de exceções.

Esse trabalho utilizou bases de dados mensais, no período de janeiro a dezembro de 2006, em valores originais de cada variável e descritos em detalhes mais a frente. Entretanto os mesmos foram padronizados, cujos valores foram usados na tentativa de se conseguir maiores precisão e significância nos resultados além da sua confidencialidade. Considerando que as empresas analisadas têm tamanhos diferentes e mercados potencialmente heterogêneos, portanto com volume de vendas bastante diferenciadas, houve necessidade de se padronizar os valores observados. Para tanto, foi adotado o processo nos quais os dados originais foram transformados em novas variáveis com média zero e desvio padrão um (LAROSE, 2006). Isso nos permite comparar diretamente o efeito relativo de cada variável independente sobre a variável dependente, facilitando sobremaneira a interpretação dos dados a serem analisados, além de reservar o caráter de confidencialidade dos dados comerciais.

Em relação às medidas de desempenho disponibilizadas, essa pesquisa desenvolveu-se com a utilização de 17 variáveis descritas anteriormente, as quais fazem parte da sistemática de acompanhamento de desempenho de vendas das empresas analisadas, relativas à composição das variáveis independentes e ainda uma variável dependente. Para aplicar o procedimento de regressão, é selecionada a Produtividade de Vendas (REFX) como a variável dependente (resposta a uma mudança nas variáveis independentes) a ser explicada pelas variáveis independentes (causa presumida de qualquer mudança na variável dependente) que representam os indicadores de desempenho das empresas de telecomunicações analisadas. As 17 variáveis a seguir descritas são incluídas como independentes: CO_1 (Venda x Mapeamento da Concorrência); EO_1 (Assertividade de Entrega e Cadastro de Contratos); EO_4 (Retorno de Contratos); EO_5 (Abertura de Ordem

de Serviço); EO_6 (Conversão de Contratos); EO_7 (Cancelamento de Vendas); FC_1 (Média de Visitas por Consultor); FC_2 (Assertividade de Agendamento); FC_3 (Visitação X Agendamento); GP_1 (Produtividade de Visita); GP_2 (Eficácia de Visitas); GP_3 (Cesta de Produtos); GP_4 (Distribuição de Vendas por Consultores); GP_5 (Produtividade de Oportunidades); IA_2 (Volume de Vendas com Descontos); IA_3 (Aprovação de Viabilidades); IA_4 (Assertividade de Analise de Viabilidade).

4.1 Modelando dados com uso de análise de regressão multivariada

Regressão é o termo utilizado para designar uma equação matemática que descreva as relações entre duas ou mais variáveis. Regressão linear é um método para se estimar o valor esperado de uma variável *Y* (variável dependente), dados os valores de algumas outras variáveis *X* (variáveis independentes). Assim, dadas duas matrizes de dados, *X* e *Y*, a finalidade da regressão é construir um modelo $Y = f(X)$. Tal modelo tenta explicar, ou predizer, as variações em *Y* dada as variações em *X*. A regressão multivariada leva em consideração as diversas variáveis preditivas simultaneamente, modelando a variável dependente com mais exatidão. Nesse trabalho, a variável dependente são as vendas efetivas e o grupo de variáveis independentes são os indicadores do desempenho de vendas. O modelo de regressão é representado pela equação 1.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \,. \qquad (1)$$

Em que Y_i – representa a variável dependente, x_{ik} ($i = 1, ..., n$) são as variáveis independentes ($k = 1, 2, ..., p$); β_i's são os coeficientes da regressão (parâmetros desconhecidos no modelo – a serem estimados); ε_i é o resíduo, variável aleatória que captura a parcela do comportamento da variável Y_i não explicada pela equação da regressão.

As etapas básicas para realizar a análise de regressão são:

- Formular o modelo geral;
- Estimar os parâmetros;
- Estimar coeficientes padronizados da regressão;
- Testes de significância do modelo;
- Determinar as relações entre variáveis;
- Verificar a exatidão das predições;
- Examinar os resíduos;
- Realizar a validação cruzada do modelo.

Um dos aspectos os mais importantes nessa análise é a seleção das variáveis independentes a ser usadas na regressão. As variáveis independentes possíveis que podem influenciar na produtividade de vendas devem ser listadas a priori, reduzindo o custo da pesquisa (LAROSE, 2006). Entretanto, objetiva-se detectar o subconjunto de variáveis, dentre todas as disponíveis, que seja mais relevante para a formação das vendas.

Os parâmetros de um modelo da regressão podem ser estimados de várias formas:

1) Mínimos quadrados, minimizando o erro quadrático médio dos resíduos;
2) Máxima verossemelhança;
3) Métodos Bayesianos;
4) Minimizando o desvio absoluto.

Os métodos 1 e 2 coincidem para um modelo com os erros normalmente distribuídos. Estimativas dos mínimos quadrados, usados nesse trabalho, são dadas por (LAROSE, 2006)

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}. \qquad (2\text{-a})$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}. \qquad (2\text{-b})$$

O estimador de mínimos quadrados, na forma matricial, é dado por $\beta = (X'.X)^{-1} (X'.Y)$, onde o apóstrofo significa transposto. Cada observação tem seu próprio resíduo, que somados produzem a soma dos erros quadráticos, uma medida total dos erros da estimação. Três somas quadráticas (SSE, soma quadrática dos erros; SSR, a soma dos quadrados da regressão; SST, a soma total dos quadrados) podem ser calculadas como segue:

$$SSE = \sum (y - \hat{y})^2 . \qquad (3\text{-a})$$

$$SSR = \sum (\hat{y} - \bar{y})^2 . \qquad (3\text{-b})$$

$$SST = \sum (y - \bar{y})^2 . \qquad (3\text{-c})$$

A estatística da regressão pode ser apresenta sucintamente com uso de tabelas da análise de variância (ANOVA). Erros médios (por exemplo, MSE e o MSR) são derivados da equação 3. Um parâmetro importante é o coeficiente de determinação múltipla, que é definida como:

$$R^2 = \frac{SSR}{SST}. \qquad (4)$$

Para a regressão múltipla, R^2 é interpretado como a proporção da variabilidade na variável alvo que é esclarecida no relacionamento linear com o conjunto de variáveis preditoras.

O uso de análise da regressão pressupõe aderência a um conjunto de suposições:

- As variáveis preditoras devem ser linearmente independentes;
- Os termos do erro devem ser normalmente distribuídos e independentes;
- A variância dos termos de erro deve ser constante;
- Deve ser utilizada uma amostra representativa da população para correta inferência;
- A distribuição da variável dependente deve ter variância aproximadamente constante, isto é, suposição do homoscedasticidade.

Os principais problemas que devem ser enfrentados em uma regressão são a multicolinearidade, heteroscedasticidade e autocorrelação.

4.2 Métodos de seleção de variáveis testados no modelo de regressão

Existem vários métodos de seleção de variáveis para construção dos modelos de regressão. Os métodos diferentes conduzem a uma variedade de modelos da regressão de um mesmo conjunto de (SPSS, 2007). Para tanto, foram utilizados nesse trabalho os seguintes métodos:

- Um modelo de regressão linear específico (ou seja, contendo as variáveis explicativas especificadas pelos especialistas de negócios da área de telecomunicações) é ajustado em primeiro lugar e em seguida são efetuados testes individuais aos coeficientes ajustados;
- Um modelo de regressão linear completo (ou seja, contendo todas as variáveis explicativas disponíveis) é ajustado em segundo lugar e em seguida são efetuados testes individuais aos coeficientes ajustados (*Enter*, no SPSS);
- Um novo modelo, *Stepwise*, no SPSS (por conter menos variáveis explicativas que o Combinatório) é ajustado, tendo-se retirado aquelas variáveis que demonstraram não contribuir para a explicação da variação da variável dependente.

A fim de comparar diversos modelos da regressão, é feito inicialmente um estudo confirmatório onde o investigador especifica o grupo completo das variáveis independentes a ser incluídas no modelo (primeiro caso), denominado de Especificação Confirmatória. Em seguida a metodologia é repetida, porém com outro método de busca por meio de uma abordagem combinatória que é um processo de busca generalizada em todas as possíveis combinações das variáveis independentes pelo método Combinatório (*Enter*). Por último, é utilizado o método de busca seqüencial para estimar a equação de regressão com um conjunto de variáveis sendo então acrescentado seletivamente ou eliminado variáveis até que uma medida de critério geral seja alcançada, sendo denominado Seqüencial (*Stepwise*). São avaliados os coeficientes ajustados e o impacto potencial das variáveis omitidas para garantir que a significância gerencial seja avaliada juntamente com a significância estatística.

No estudo confirmatório, um conjunto de variáveis estabelecidas pelo especialista, deve ser confirmada se há correlação com a variável dependente. No método *Enter* um bloco (ou todas) são selecionadas e incorporadas ao modelo em uma única etapa. No método *Stepwise*, em cada etapa, a variável independente ainda ausente da equação são incorporadas caso possuam valores baixos da estatística *F*. As variáveis já na equação da regressão são removidas se sua probabilidade de *F* se tornar suficientemente grande. O método conclui quando não mais há variável elegível para a inclusão ou a remoção.

Outros métodos no SPSS disponíveis incluem: *Remove*, que permite remover blocos de variáveis em uma única etapa; *Backward Elimination*, o qual todas as variáveis são incorporadas na equação e então removidas sequencialmente (baseado nas menores correlações parciais com a variável dependente); e *Forward Selection*, que incorpora sequencialmente variáveis no modelo. Nesse último, a primeira variável considerada é a que possui maior (em módulo) correlação com a variável dependente. Essa variável será incorporada na equação se satisfizer a um critério (ex. limiar) para a entrada. O processo continua com a segunda variável com maior correlação parcial (em módulo) e assim em diante. O procedimento encerra quando não há mais variáveis que satisfaçam o critério. Todas as variáveis devem passar o critério da tolerância para entrarem na equação de regressão, indiferente do método especificado. O nível de tolerância usado foi 0.05.

4.3 Análise de Variância

A análise da variância permite decompor a variação total observada na variável dependente em variação explicada pela função de regressão (Soma dos Quadrados da Regressão ou *SSR*) e variação não explicada pela função de regressão (Soma dos Quadrados dos Erros ou *SSE*). A partir desses dois valores é possível obter a Média da Soma dos Quadrados da Regressão (*MSR*) e a Média da Soma dos Quadrados dos Erros (*MSE*).

O cálculo do *MSE* e do *MSR* permite efetuar um teste de significância global do modelo, utilizando à estatística *F* (*F-test*).

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$ (ausência de efeito);

$H_1 : \beta_1 \neq 0 \vee \beta_2 \neq 0 \vee \beta_3 \neq 0 \vee \beta_4 \neq 0 \vee \beta_5 \neq 0 \vee \beta_6 \neq 0$ (presença de efeito).

A hipótese H_0 significa que a regressão não é significativa, ou seja: que a equação de regressão não explica a variação na variável resposta e que não existe relação linear entre a variável dependente e o conjunto de variáveis independentes utilizadas.

A estatística *F** é calculada a partir da razão entre *MSR* e *MSE*. Quanto o valor de *F** é elevado, significa que uma grande parte da variação dos valores observados para a variável endógena é explicada pela reta de regressão. Caso contrário, quando *F** é reduzido, a reta de regressão apenas explica uma pequena parte da variação da variável endógena.

O critério de aceitação da hipótese nula resulta da comparação entre *F** e o valor da estatística *F* para *p-1* e *n-p* graus de liberdade (*n* é o número de observações e *p* o número de coeficientes do modelo) (NORUSIS, 2004). Assim, se $F^* \leq F(1-\alpha;\ p-1\ ;\ n-p\ )$, deve-se aceitar a hipótese nula, caso contrário, rejeita-se. A existência de uma relação de regressão, por si só, não garante que predições úteis podem ser feitas usando esse modelo. Esse teste é apenas uma etapa na verificação de aceitação do modelo.

4.4 Testes Individuais dos Parâmetros

Os coeficientes de regressão ajustados são usados para calcular os valores previstos para cada observação e para expressar a variação esperada na variável dependente para cada variação unitária nas variáveis independentes. Nesse trabalho, o objetivo é saber quais indicadores de desempenho (variáveis independentes) têm maior efeito na previsão da produtividade de vendas (variável dependente). Em alguns casos, os coeficientes de regressão não fornecem literalmente essa informação. Para resolver esse problema de explicação é usado um coeficiente de regressão modificado chamado de coeficiente beta os quais são os coeficientes resultantes de dados padronizados que eliminam o problema de lidar com diferentes unidades de medidas, refletindo assim o impacto relativo sobre a variável dependente de uma mudança em um desvio padrão de qualquer variável (NORUSIS, 2004).

Os testes individuais dos parâmetros β_i são realizados utilizando à estatística *t* de *Student*:

$H_0 : \beta_i = 0$ (ausência de efeito);

$H_1 : \beta_i \neq 0$ (presença de efeito).

$$t^* = \frac{b_i}{s(b_i)} \cap t_{(n-p)} \qquad i = 1,..., p$$

em que b_i é o valor ajustado para o parâmetro β_i, e $s(b_i)$ o valor ajustado para o desvio padrão de b_i.

Se $|t^*| \leq t\left(1-\frac{\alpha}{2}; n-p\right)$, então se deve aceitar a hipótese nula. Caso contrário, deve-se rejeitá-la.

5 - AVALIAÇÃO DOS RESULTADOS OBTIDOS

O estudo de viabilidade de uso de análise de regressão linear múltipla situa-se no domínio comercial do setor corporativo de telecomunicações, utilizando-se medição de desempenho para a identificação de relações entre parâmetros de vendas e de indicadores de desempenho. Dentro desse contexto, as hipóteses analisadas no presente trabalho são:

1. *H0*: Não existe relação entre as características da produtividade de vendas e os indicadores de desempenho de vendas.

2. *H1*: Existe relação entre as características da produtividade de vendas e as variáveis de desempenho de vendas das empresas

A variável dependente escolhida foi produtividade de vendas e um conjunto de variáveis independentes relacionados aos indicadores de desempenho de vendas. Foram avaliados o modelo de regressão e a precisão preditiva das variáveis independentes, após o uso dos métodos de seleção de variáveis. Outros fatores levados em conta incluem a significância estatística do modelo geral na previsão da variável dependente e a busca por observações com influência indevida nos resultados. Assumiu-se que os termos de erro ε_i são independentes e seguem uma distribuição $N(0, \sigma^2)$.

Foi utilizado o programa *SPSS* para estimar os parâmetros do modelo, por meio do método dos mínimos quadrados. São apresentados os coeficientes do modelo ajustados, respectivos desvios padrão, resultados dos testes individuais (estatística *t* e nível de significância), intervalo de confiança, correlações e estatísticas de colinearidade.

Objetiva-se avaliar os diversos métodos seleção de variáveis de regressão utilizados o Método de Especificação Confirmatória, o Método Combinatório (*Enter* no SPSS) e o Método Seqüencial (no SPSS é denominado de *Stepwise*) por meio da estimação dos parâmetros, do coeficiente de determinação múltiplo, da análise de variância e dos testes individuais dos parâmetros.

O teste de hipótese é uma regra usada para decidir se uma hipótese estatística deve ser rejeitada ou não, isto é, se uma hipótese sobre determinada característica da população é ou não apoiada pela evidência obtida dos dados amostrais. Em análise de regressão os testes de hipóteses necessários são: teste de hipótese para a significância do modelo, teste de hipótese para o parâmetro βi e o teste de hipótese para um subconjunto de parâmetros (NETER & WASSERMAN, 1974).

A tabela 1 sintetiza resultados da aplicação dos três métodos de seleção de variáveis testados. Analisando-se as informações disponibilizadas pelo coeficiente de determinação múltiplo, Teste *F* e o teste estatístico *t,* pode ser verificado que:

- O Método Combinatório comprova a não validação da hipótese H_0 tendo em vista que os coeficientes β_1, β_2, β_{10}, β_{11}, β_{14} e β_{15} obtiveram valores de significância *sig* < 0,05 sendo, portanto significantes, caracterizando o relacionamento dos indicadores de desempenho de vendas associados a esses coeficientes com a produtividade de vendas.
- O Método Confirmatório comprova a validação da hipótese H_1 tendo em vista que o coeficiente β_3 obteve um valor de significância *sig* > 0,05 não sendo, portanto estatisticamente significante. Essa exclusão da variável *FC1*, a qual está associada ao coeficiente β_3, caracteriza a não confirmação das variáveis, especificadas na equação de regressão linear múltipla, definidas pelos especialistas de vendas de telecomunicações.
- O Método Seqüencial comprova também a validação da hipótese H_1 tendo em vista que os coeficientes β_1, β_2, β_3, β_4, β_5 e β_6 obtiveram valores de significância *sig* < 0,05 sendo, portanto significativos, caracterizando o relacionamento dos indicadores de desempenho de vendas associados a esses coeficientes com a produtividade de vendas. Porém essa relação é diferente da especificada pelos especialistas de vendas em serviços de telecomunicações.

Tabela 1 – Comparativo dos métodos testados.

	Métodos Testados		
	Especificação Combinatória (Enter)	Especificação Confirmatória	Especificação Sequencial (Stepwise)
Parâmetros do Modelo	**REFXi = 0,289 CO1i + 0,204 EO1i + 0,010EO4i – 0,028 EO5i – 0,024 EO6i – 0,065 EO7i + 0,030 FC1i + 0,067 FC2i – 0,085 FC3i + 0,184 GP1i + 0,161 GP2i + 0,010 GP3i + 0,062 GP4i + 0,234 GP5i + 0,132 IA2i + 0,026 IA3i – 0,059 IA4i**	**REFXi = 0,293 CO1i + 0,198 EO1i + 0,060 FC1i + 0,166 GP1i + 0,203 GP2i + 0,251 GP5i + 0,131 IA2i**	**REFXi = 0,299 CO1i + 0,207 EO1i + 0,168 GP1i + 0,196 GP2i + 0,254 GP5i + 0,137 IA2i**
Coeficiente de Determinação Múltiplo	0,501	0,481	0,478
Teste F	13,807>>1,70	32,368>>2,05	37,406>>2,10
Teste Estatístico t	**||0,188| < |0,196| < |0,493| < |0,513| < |0,565| < |0,587| < |1,006| < |1,106| < |1,189| < |1,276| < |1,588| < 2,241 < |2,679| < |2,796| < |3,536| < |3,945| < |4,541| < |5,605|**	**|1,263| < 2,241< |2,749| < |3,390| < |3,984| < |4,221| < |5,088| < |5,908**	**| *t* i* | ≥ 2,241 (i = 1,...,6)**
Variáveis Excluídas	Nenhuma	FC1	EO4, EO5, EO6, EO7, FC1, FC2, FC3, GP3, GP4, IA3 e IA4
Quantidade de Variáveis na Equação	17	7	6
Validação da Hipótese H0	Não	Não	Não
Validação da Hipótese H1	Sim	Sim	Sim

Fonte: Elaborada pelo autor a partir dos dados do SPSS.

Como conclusão, temos que o modelo *Stepwise,* que escolhe CO_1, EO_1, GP_1, GP_2, GP_5 e IA_2 como regressores, é o mais apropriado para ser utilizado pela gerência na explicação da Produtividade de Vendas com os indicadores de desempenho. O mesmo possui características gerenciais adequadas para a finalidade que se propôs (equação 5).

$$REFX_i = 0{,}299 \cdot CO_1 + 0{,}207 \cdot EO_1 + 0{,}168 \cdot GP_1 + 0{,}196 \cdot GP_2 + 0{,}254 \cdot GP_5 + 0{,}137 \cdot IA_2. \quad (5)$$

As vendas são afetadas positivamente por esses indicadores de desempenho, pois para todos eles *t* foi maior que *t* (97,5%;245) ≅ 2,241. Vendas x Mapeamento da Concorrência foi escolhido primeiramente porque é o regressor com maior correlação com vendas. Os regressores restantes são analisados qual o mais apropriado para a inclusão na etapa seguinte. Para escolher a melhor variável a adicionar ao modelo, olha-se na correlação parcial, que é a correlação linear entre o regressor proposto e a variável dependente após ter sido removido o efeito do modelo atual (LAROSE, 2006). Assim, Produtividade de Oportunidade foi escolhido em seguida porque tem a correlação parcial mais elevada.

A. Variabilidade da Produtividade de Vendas em Função das Correlações

O coeficiente de correlação semiparcial é a correlação entre a variável dependente produtividade de vendas e um regressor parcializado dos restantes. O quadrado dessa quantidade nos fornece a proporção da variabilidade da produtividade de vendas explicada exclusivamente pelo regressor. O coeficiente de correlação parcial é a correlação entre duas variáveis quando ambas foram parcializadas de variáveis terceiras. O quadrado dessa quantidade dá-nos a proporção da variabilidade da produtividade de vendas não associada a um regressor x_i que está associada ao outro regressor x_k (LAROSE, 2006). Em outras palavras, responde a questão: "quanto da variância da produtividade de vendas que não é estimada pelas outras variáveis independentes na equação é estimada por essa variável". Pode ser dimensionada essa variabilidade (tabela 2) usando os dados de correlações semiparcial e parcial obtidos.

Tabela 2 – Produtividade de vendas em função das correlações do método seqüencial.

Indicador de Desempenho	**Correlações**					
	Semiparcial			**Parcial**		
	sr	**sr²**	**%**	**pr**	**pr²**	**%**
CO1 Venda x Mapeamento da Concorrência	0,279	0,0778	7,8	0,360	0,130	13,0
GP5 Produtividade de Oportunidades	0,238	0,0566	5,7	0,312	0,097	9,7
GP1 Produtividade de Visita	0,159	0,0253	2,5	0,214	0,046	4,6
EO1 Assertividade de Entrega e Cadastro de Contratos	0,193	0,0372	3,7	0,258	0,067	6,7
GP2 Eficácia de Visitas	0,189	0,0357	3,5	0,254	0,065	6,5
IA2 Volume de Vendas Com Descontos	0,133	0,0177	1,8	0,182	0,033	3,3

Fonte: Elaborado pelo autor a partir dos dados do SPSS.

B. Teste F Generalizado do Modelo Stepwise em Relação ao Modelo Combinatório

Ao construir o modelo *Stepwise* a partir do modelo Combinatório, retirando onze variáveis independentes, verifica-se que a variância total da variável endógena do modelo que é explicada pela função de regressão diminui. Ou seja, a soma do quadrado dos erros no modelo *Stepwise* é sempre superior a soma do quadrado dos erros no modelo Combinatório, ou $SSE_s > SSE_c$. Após o teste *t* sugerir as variáveis independentes a serem usadas na equação, é importante examinar se a variável dependente pode ser explicada pelas variáveis sugeridas tão adequadamente quanto por todas as variáveis. Para isto, testam-se as hipóteses:

$$H_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_{q+p} = 0 \text{ , } q<p$$
$$H_1 : \beta_k \neq 0 \text{ , para algum } k=q+1, \ldots, p$$

em que *q* representa os coeficientes não usados na equação.

O teste *F* generalizado pretende determinar se o acréscimo de *SSE* resultante da redução do modelo é ou não significativo. Se não for significativo, então de fato as variáveis excluídas não contribuíam significativamente para diminuir a variação não explicada pelo modelo. Para a realização do teste *F* generalizado, é utilizada a seguinte estatística:

$$F = \frac{SSE_s - SSE_c}{GLE_s - GLE_c} \Big/ \frac{SSE_c}{GLE_c} . \quad (6)$$

Em que GLE_s e GLE_c representam os graus de liberdade associados ao erro no modelo Stepwise e no modelo Combinatório, respectivamente. Esse valor é depois comparado com o valor de $F(1-\alpha ; GLE_s-GLE_c ; GLE_c)$, e se for inferior, então deve ser aceito a hipótese de que não existem diferenças significativas entre SSE_s e SSE_c. A partir das tabelas ANOVA para os métodos Combinatório e *Stepwise* obtidas por intermédio do *SPSS*, pode-se calcular o valor do teste F generalizado.

Se $F \leq F(p-q; n-p-1)$, o teste não rejeita H_0; caso contrário o teste rejeita H_0 em favor de H_1. A aceitação de H_0 indica que a variação da variável dependente é tão adequadamente explicada como o conjunto de todas as variáveis independentes (MICKEY & CLARCK, 2005). Da tabela da estatística *F* pode ser verificado que $F_{(95\% ; 11 ; 245)} \cong 1{,}85$. Como 0,966 < 1,85, então não existem diferenças significativas entre os dois modelos, ou seja, as variáveis retiradas do modelo completo não contribuíam significativamente para a redução da variabilidade não explicada (medida pelo *SSE*).

C. *Interpretação dos Coeficientes do Modelo Seqüencial*

Os valores dos coeficientes das variáveis podem ser interpretados da seguinte forma:

a) O valor do coeficiente ajustado para a variável CO_1 (0,299) significa que mantendo todas as outras variáveis constantes, para cada venda realizada com mapeamento da concorrência em relação ao total de vendas, a produtividade aumenta, em média,

aproximadamente R$ 299,00 sempre que o volume de negócios aumentar R$ 1.000,00 normatizados. Por outro lado, 7,8% da variabilidade das vendas devem-se exclusivamente a CO_1 quando as vendas realizadas estão mapeadas da concorrência e que as não explicadas pelas outras variáveis independentes, CO_1 explicam 13%;

b) O valor do coeficiente ajustado para a variável GP_5 (0,254) significa que mantendo todas as outras variáveis constantes, para cada venda realizada com mapeamento de oportunidades em relação ao total de visitas com oportunidades identificadas, à produtividade aumenta, em média, aproximadamente R$ 254,00 sempre que o volume de negócios aumentar R$ 1.000,00 normatizados. Por outro lado, 5,7% da variabilidade das vendas devem-se exclusivamente a GP_5 quando a equipe de consultores fecha negócios nas oportunidades prospectadas e que as não explicadas pelas outras variáveis independentes, GP_5 explicam 9,7%;

c) O valor do coeficiente ajustado para a variável GP_1 (0,168) significa que mantendo todas as outras variáveis constantes, para cada visita realizada com oportunidades identificadas em relação ao total de visitas, a produtividade aumenta, em média, aproximadamente R$ 168,00 sempre que o volume de negócios aumentar R$ 1.000,00 normatizados. Por outro lado, 2,5% da variabilidade das vendas devem-se exclusivamente a GP_1 quando a equipe de consultores identifica oportunidades nas visitas realizadas e que as não explicadas pelas outras variáveis independentes, GP_1 explicam 4,6%;

d) O valor do coeficiente ajustado para a variável EO_1 (0,207) significa que mantendo todas as outras variáveis constantes, para cada contrato cadastrado no prazo em relação ao total de contratos cadastrados, a produtividade aumenta, em média, aproximadamente R$ 207,00 sempre que o volume de negócios aumentar R$ 1.000,00 normatizados. Por outro lado, 3,7% da variabilidade das vendas devem-se exclusivamente a EO_1 quando os contratos dos serviços comercializados são entregues e cadastrados no prazo e que as não explicadas pelas outras variáveis independentes, EO_1 explica 6,7%;

e) O valor do coeficiente ajustado para a variável GP_2 (0,196) significa que mantendo todas as outras variáveis constantes, para cada contrato fechado em relação ao total de visitas realizadas, a produtividade aumenta, em média, aproximadamente R$ 196,00 sempre que o volume de negócios aumentar R$ 1.000,00 normatizados. Por outro lado, 3,5% da variabilidade das vendas devem-se exclusivamente a GP_2 quando os contratos dos serviços comercializados são coerentes com as visitas realizadas e que as não explicadas pelas outras variáveis independentes, GP_2 explicam 6,5%;

f) Finalmente, o valor do coeficiente ajustado para a variável IA_2 (0,137) significa que mantendo todas as outras variáveis constantes, para cada venda realizadas com customização em relação ao total de vendas, a produtividade aumenta, em média, aproximadamente R$ 137,00 sempre que o volume de negócios aumentar R$ 1.000,00 normatizados. Por outro lado, 1,8% da variabilidade das vendas devem-se exclusivamente a IA_2 quando os serviços comercializados são menores que 30% de desconto e que as não explicadas pelas outras variáveis independentes, IA_2 explica 3,3%.

D. *Análise dos Pressupostos do Modelo Seqüencial*

O método de regressão linear múltipla, apesar de não ser muito complexo de implementar, está estruturado num conjunto de hipóteses fundamentais que nem sempre se verificam na prática, quais sejam: não multicolinearidade, a homoscedasticidade, normalidade da distribuição dos termos de erros e a ausência de autocorrelação nos erros. Nessa secção, analisa-se a natureza e as conseqüências práticas da verificação ou não desses pressupostos.

1) Multicolinearidade

Comparando o modelo Combinatório com o *Stepwise*, se verifica que a inclusão ou exclusão de variáveis não altera significativamente os coeficientes ajustados b_i para a função de regressão. Por si só, essa observação é um sinal da não existência de multicolinearidade, uma vez que quando as variáveis independentes do modelo estão correlacionadas, os coeficientes ajustados variam consideravelmente como conseqüência da introdução ou exclusão de variáveis no modelo.

Por outro lado, fazendo-se análise da matriz de correlação entre as variáveis do modelo, pode ser verificado que, entre as variáveis independentes, não existem valores superiores a 0,5, conforme dados obtidos a partir do *SPSS* (NORUSIS, 2004). Ou seja, não existindo nenhum valor $r_{X_i X_j} \geq 0{,}5$, pode ser concluído pela inexistência de multicolinearidade. O *SPSS* fornece ainda outro indicador, mais formal, da existência de multicolinearidade: o Fator de Inflação da Variância (*FIV*). O *FIV* é uma medida do grau em que a correlação entre as variáveis independentes faz inflacionar a variância associada à distribuição dos coeficientes ajustados da regressão.

$$FIV_i = \left(1 - R_i^2\right)^{-1} \quad i = 1,2,..., p-1. \quad (7)$$

Em que R_i^2 corresponde ao coeficiente de determinação múltipla de um modelo de regressão linear e a variável X_i é dependente das restantes variáveis X_j. Uma maneira adequada para verificar a colinearidade de duas ou mais variáveis é por meio do valor de Tolerância ou do Fator de Inflação de Variância (*variance inflation factor VIF*) que é o seu inverso. Essas medidas dizem o grau em que cada variável independente é explicada pelas demais variáveis independentes. Pequenos valores de Tolerância ou elevados de *FIV* denotam colinearidade elevada, sendo usual a utilização de valores de referência para a Tolerância na ordem de 0,10, o que corresponde a um valor de *FIV* acima de dez. Os dados obtidos mostram que os valores do *FIV* são bastante próximos de um, pelo que se pode concluir pela não existência de Multicolinearidade.

Outra maneira que ajuda na determinação da ocorrência de problemas com colinearidade é por meio dos autovalores que possibilitam uma estimação do número de variáveis independentes. Instabilidades podem ocorrer quando muitos autovalores são próximos a zero. Nesse caso as variáveis são altamente intercorrelacionadas e as pequenas mudanças nos valores dos dados podem conduzir a grandes mudanças nas estimativas dos coeficientes (NORUSIS, 2004).

Os Índices de Condições são as raízes quadradas das relações entre autovalores sucessivos, após ordenamento do maior para o menor. Um índice de condições superior a 15 indica um problema possível e um índice superior a 30 sugere um problema sério em relação à colinearidade. As proporções da variância são as proporções da variância da estimativa esclarecida por cada componente principal associado com cada um dos autovalores. Colinearidade é um problema quando um componente associado com um índice de condições elevado contribui substancialmente para variância de duas ou mais variáveis.

O diagnóstico de colinearidade confirma que não existe nenhum problema no modelo encontrado. Os autovalores encontrados são muito maiores que zero, indicando que os regressores não são intercorrelacionados e que apenas na ocorrência de grandes mudanças nos valores dos dados podem conduzir a pequenas mudanças nas estimativas dos coeficientes.

2) Heteroscedasticidade

A presença de variâncias desiguais (heteroscedasticidade) é uma das violações mais comuns de suposições em regressão. Esse diagnóstico pode ser feito avaliando o comportamento geral dos resíduos utilizando um gráfico, contendo no eixo vertical os resíduos padronizados e no eixo horizontal os valores previstos padronizados. Também devem ser construído um diagrama de dispersão para visualizar a relação entre a variável dependente (Produtividade de Vendas) e cada um dos regressores individualmente. Pode-se visualizar também o gráfico de dispersão que apresente os valores observados da variável dependente com os valores preditos padronizados.

Pela configuração da dispersão de valores dos resíduos padronizados e resíduos do método seqüencial, pode-se concluir pela existência de homocedasticidade, dado não haver um padrão de variação dos termos de erro ajustados pelo modelo relativamente aos valores ajustados da variável dependente. Também se pode concluir pela não existência de uma relação sistemática entre os valores ajustados do termo de erro, e as variáveis independentes do modelo. Mais uma vez, essas observações corroboram a hipótese da não existência de heteroscedasticidade, ou seja, os gráficos de ajustamento dos resíduos não apresentam nenhum tipo de tendência que faça suspeitar da validade dos pressupostos do modelo de regressão, pois os resíduos distribuem-se segundo uma faixa horizontal em torno do zero, sem denotar qualquer padrão de distribuição.

A independência dos termos de erro pode ser identificada fazendo o gráfico de resíduos em relação a qualquer variável seqüencial possível. Se os resíduos forem independentes, o padrão é aleatório e semelhante ao gráfico nulo de resíduos, pois é assumido em regressão que cada valor previsto é independente, ou seja, não está relacionado com qualquer outra previsão. A homoscedasticidade é uma suposição relacionada primariamente a relações de dependência entre variáveis, referindo-se à suposição de que a variável dependente exibe níveis iguais de variância ao longo do domínio da variável preditora.

3) Normalidade

O diagnóstico mais usual para verificação de normalidade da distribuição dos termos de erros é um histograma de resíduos, com uma verificação visual para uma distribuição que se aproxima da normal (ver figura 1). Um histograma é uma representação gráfica de uma única variável que representa a freqüência (valores dos dados) dentro da categoria de

dados. O histograma traduz a distribuição de freqüências, sendo possível analisar a simetria e o achatamento da amostra. Esse gráfico sintetiza a estrutura da população de onde foi retirada a amostra. No caso de distribuições a área total do gráfico deve ser unitária.

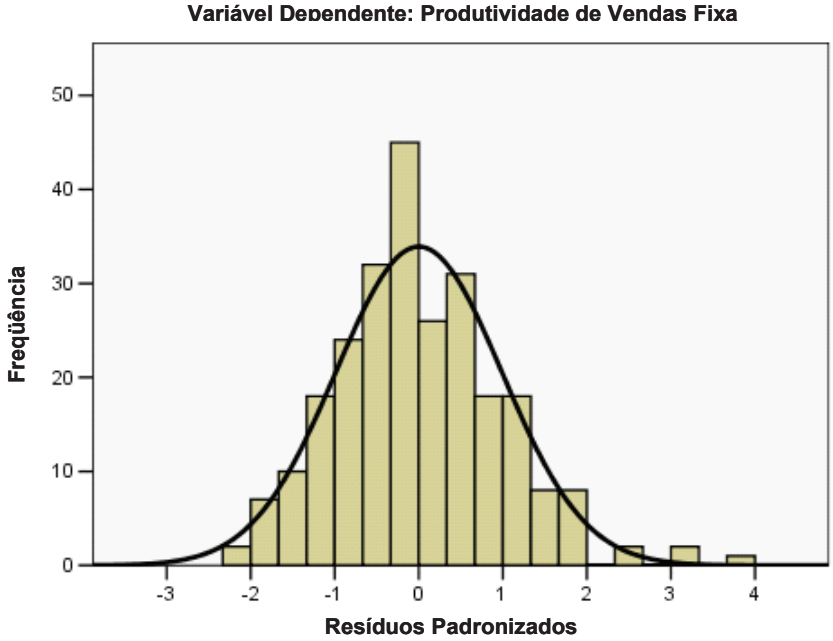

Figura 1 – Histograma de resíduos da variável dependente do método seqüencial.

Fonte: Relatório do SPSS.

Uma outra abordagem utilizada é o compara a distribuição cumulativa de dados com a distribuição cumulativa de uma distribuição normal. A distribuição normal forma uma reta diagonal. Se uma distribuição é normal, a linha que representa os dados segue muito próximo à diagonal. Os gráficos de probabilidades (*PP-plot*: *Probability Plots*) visualizam graficamente o ajustamento de uma variável a uma função de distribuição de probabilidades. Esse tipo de gráfico representa no eixo horizontal as freqüências relativas acumuladas observadas na amostra (*observed cummulative probability*) e, no eixo vertical, a função de distribuição de probabilidades esperada (*expected cummulative probability*). A diagonal do gráfico representa um ajustamento perfeito da amostra à função de distribuição de probabilidades. Quanto mais os pontos se afastam da diagonal, ou se distribuem segundo um determinado padrão, menor é o ajustamento da amostra à distribuição teórica.

No histograma apresentado na figura 1, com o ajustamento à distribuição normal, pode ser verificado que existe uma distribuição consistente dos resíduos em relação à distribuição teórica, ou seja, segue aproximadamente a forma da curva normal. Existe apenas uma elevação um pouco maior, nomeadamente na zona central da distribuição, sendo aceitável perto da curva normal, porém com a existência de pontos residuais positivos relativamente grandes.

Na figura 2 nota-se uma tendência de distribuição uniforme, ou seja, os pontos encontram-se bastante próximos de uma reta, sem desvios substanciais ou sistemáticos. Não há razão para duvidar da normalidade dos erros, sendo os resíduos considerados representativos de uma distribuição normal. Nem o histograma nem o gráfico *PP-plot* indicam que a suposição da Normalidade foi violada.

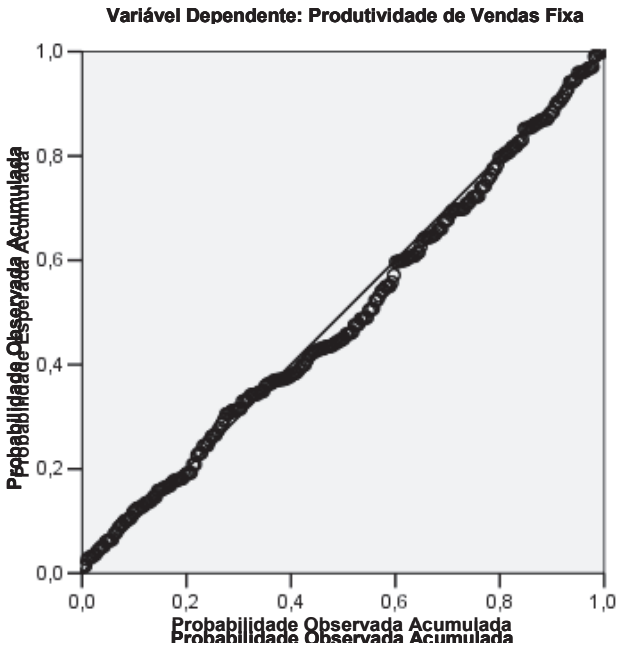

Figura 2– Probabilidade normal PP-plot dos resíduos do método seqüencial.
Fonte: Relatório do SPSS.

4) Autocorrelação dos Erros

A Auto-correlação ocorre quando os termos de erro não são independentes, ou seja, $Cov(\varepsilon_i, \varepsilon_j) = 0$. A autocorrelação entre os resíduos pode ser detectada pelo método gráfico ou por meio do teste *d* de Durbin-Watson. Na analise gráfica por meio do gráfico de resíduos (ε_i) *versus* valores unitários ajustados pelo modelo de regressão (y), observou-se que os resíduos estavam distribuídos aleatoriamente em torno da reta não apresentando nenhum padrão definido. Dessa forma, houve a necessidade de se aplicar o teste *d* de Durbin-Watson para se detectar a existência de autocorrelação positiva ou negativa. O teste de autocorrelação Durbin-Watson é obtido a partir dos resíduos do modelo no qual são estabelecidos os valores críticos do limite inferior (d_i) e do limite superior (d_u) que é verificado por meio de tabelas em função de p (regressores), n (número de observações) e α (nível de significância), sendo que: para $d < 2$ temos: se $d < d_i$ rejeita-se H_o, ou seja, que os resíduos aleatórios não são autocorrelacionados e se aceita a autocorrelação, se $d_i < d < d_u$ então o teste é inconcluso e se $d > d_u$ se aceita H_o; para d > 2 temos: se $d < 4 - d_u$ se aceita H_o e se aceita a autocorrelação, se $(4 - d_u) < d < (4 - d_i)$ então o teste é inconcluso e se $d > (4 - d_i)$ rejeita-se H_o.

A existência de autocorrelação foi testada recorrendo ao teste de Durbin-Watson ($p = 6$, $n = 252$ e $\alpha = 5\%$). No modelo gerado a estatística d estimada foi $d = 1.801 < 2$ e $d_u = 1.757$ enquadrando-se no intervalo de valores associados à aceitação da hipótese nula (então se aceita H_o). Por conseqüência, é de se admitir que a eficiência dos estimadores obtidos pelos métodos dos mínimos quadrados, dada à inexistência de autocorrelação entre eles. Assim, não há evidência de relação de dependência dos valores dos resíduos aleatórios.

5) Análise dos Pontos Influentes

É comum, em trabalhos experimentais, situações em que, ao obter ou analisar um conjunto de dados, se depare com um ou mais valores que aparentemente diferem razoavelmente dos outros. Esses valores produzem, dependendo da amplitude de sua dispersão em relação aos valores esperados, conclusões errôneas e distorções nos parâmetros obtidos nos modelos. As causas podem ser variadas, como erros humanos ou instrumentais. Da sua identificação depende muitas vezes a validade das conclusões que são obtidas. A média e o desvio padrão dependem da remoção ou não desses valores, e uma vez que a discussão sobre a acurácia e precisão dos dados depende desses parâmetros, torna-se evidente o cuidado a ter em relação a sua eliminação ou não, devendo sempre ser fundamentada a opção tomada. A eliminação desses valores pode igualmente incorrer num erro, como, por exemplo, sobreestimar a precisão dos dados, ou aceitar um modelo que não é válido, sendo esse pressuposto a base da classificação dos valores discrepantes (*outliers*) (MICKEY & CLARCK, 2005).

O tipo de gráfico *boxplot* permite visualizar informações importantes sobre a forma dos dados, sendo bastante útil na comparação de várias amostras. A representação de *boxplots* paralelos pode facilitar a comparação de amostras ou de coleções de dados. Essa representação permite formular conjecturas acerca das semelhanças ou diferenças com respeito às medidas de localização e de dispersão mencionadas (ver figura 3).

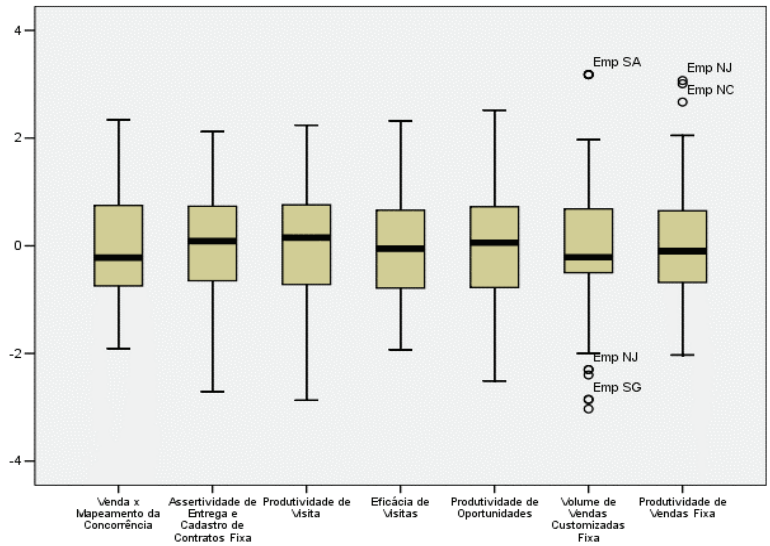

Figura 3 – *Boxplot* do método seqüencial.

Fonte: Relatório do SPSS.

Eventuais assimetrias podem ser facilmente visualizadas por meio do deslocamento da mediana para um dos lados da caixa e por meio de uma diferença significativa no comprimento da caixa. Também a presença de *outliers* apenas para um dos lados da caixa pode ser indicativa de assimetria. Na figura 3 podem ser observadas pequenas assimetrias tanto negativas quanto positivas (concentração de observações elevadas). Nota-se, por exemplo, na primeira e na sexta barra que a mediana está mais próxima do primeiro quartil do que do terceiro. Dentro dos parâmetros aceitáveis existem cinco *outliers* na variável independente *Volume de Vendas com Descontos*, sendo dois da empresa *NJ* e dois da

empresa *SG* para além da cerca inferior e um da empresa *SA* para além da cerca superior. Ocorrem três valores discrepantes na variável dependente *Produtividade de Vendas*, sendo todos além da cerca superior com duas participações da empresa *NJ* e um da empresa *NC*. Também foi presenciado cinco outros *outliers* na variável independente *Volume de Vendas com Descontos* tendo sido causado pela competitividade de preço provocado pela concorrência. Os valores observados da variável dependente são indicados, onde se pode também ver o valor predito e o valor residual para cada caso. Identificando na tabela dos dados, os casos observados acima ocorreram na empresa *NB* no mês de Dez/06 (caso 24), na empresa *NJ* no mês de Abr/06 (caso 112) e na empresa *SDT* no mês de Mar/06 (caso 243), sendo valores válidos, porém fora da curva.

6 – CONCLUSÕES E RECOMENDAÇÕES

Esse trabalho consistiu em utilizar técnicas de mineração de dados (MD) no tratamento de um banco de dados histórico proveniente da medição de desempenho organizacional na área de vendas corporativas. A proposta foi a de estabelecer os relacionamentos entre as medidas de desempenho já existentes e descobrir novos relacionamentos entre elas.

Teve como motivação o fato do sistema de gestão das empresas e, conseqüentemente, o sistema de medição de desempenho, precisar refletir as mudanças ocorridas no ambiente industrial nas últimas décadas. Um desses principais reflexos foi a necessidade do sistema de medição de desempenho se tornar multidimensional e não mais somente voltado para a avaliação dos aspectos financeiros e de produtividade. Por isso, torna-se mais complexo o entendimento de como os relacionamentos entre as medidas de desempenho afetam as decisões organizacionais.

Assim, foi delineado um método baseado no emprego de regressão linear múltipla para analisar o efeito que os indicadores de desempenho de vendas poderiam ter sobre o desempenho da produtividade de vendas. Para tanto se utilizou equações de regressão cujos parâmetros de desempenho pudessem ser favoráveis à formação de uma equação de produtividade de vendas. Mediante análises estatísticas e comerciais criteriosas, as equações foram definidas, sendo os seus respectivos coeficientes de determinação ajustados. Foram realizados testes de hipóteses dos principais parâmetros, visando à validação ou não dos modelos de regressão e a análise da qualidade de seus ajustes.

A pesquisa permitiu verificar que existe relação entre as características da produtividade de vendas e as variáveis de desempenho. Constatou-se que a relação entre a produtividade de vendas e os indicadores de desempenho acontece diferente do modelo atualmente utilizado. A variável independente *Média de Visitas por Consultor*, por exemplo, foi excluída do modelo gerado por não ter sido considerada estatisticamente significante para a composição da variável dependente Produtividade de Vendas.

Os resultados mostram que, dos dezessete indicadores de desempenho de vendas utilizados pelo modelo atual em execução, seis desses indicadores foram significativos na explicação da produtividade de venda baseada no tratamento dos dados históricos, isto é, no modelo gerado pela pesquisa. Os coeficientes de todas as variáveis foram positivos, indicando que o melhor desempenho de um dos fatores significativos conduz a um maior volume de

negócio realizado. O modelo proposto revelou bom ajuste e com boa capacidade de explicar o comportamento do volume de vendas realizado.

A variável independente *Vendas versus Mapeamento da Concorrência* foi, em relação a produtividade de vendas, a que apresentou os resultados mais importantes da pesquisa, para o conjunto das amostras utilizadas. Essa variável mostrou uma correlação positiva significativa ao nível de 5% (valor da probabilidade associada à estatística *t*). Os resultados foram consistentes em todos os casos em que existe significância, apresentando uma forte evidência de que a empresa desse setor que atua com foco no mapeamento da concorrência e na produtividade de oportunidades poderá ter melhor desempenho comercial.

Os resultados obtidos permitem sugerir que as empresas devem adotar um planejamento de vendas com ênfase nos seguintes principais indicadores de desempenho: Vendas versus Mapeamento da Concorrência, Produtividade de Oportunidades, Conversão de Contratos, Eficácia de Visitas, Produtividade de Visitas e Volume de Vendas com Descontos. Também, os parâmetros individuais do modelo, validados pela hipótese referenciada, caracterizam que o mapeamento da concorrência deve ser mais eficaz do que realizar vendas com preços com aplicação da política de descontos por alçada em relação aos valores nominais (o que caracteriza o indicador de desempenho *Volume de Vendas com Descontos*).

Novos estudos podem ser realizados, com o objetivo de melhorar a análise sobre a relação entre os indicadores de desempenho e o desempenho comercial das empresas. Sugere-se utilizar testes estatísticos alternativos, por exemplo, um modelo de análise conjunta. Ou, ainda, incluir outras empresas ou utilizar um horizonte temporal diferente do utilizado nessa pesquisa. Recomenda-se ainda a aplicação experimental de outros algoritmos para mineração dos dados selecionados para avaliação de desempenho e compatibilidade com o objetivo da aplicação, dando continuidade à pesquisa.

8 - REFERÊNCIAS

ADRIAANS, Pieter; Zanting, Dolf; "*Data Mining*", Addison-Wesley, England, 1996.

ARNETT, Dennis B; Menon, Anil, Wilcox, James B.; "*Using Competitive Intelligence: Antecedents and Consequences*", Competitive Intelligence Review, Vol. 11(3), 2000.

BITITCI, U. S; Carrie, A. S.; McDevitt, L. *Integrated performance measurement systems: a development guide*. International Journal of Operations & Production Management, V17, No 5. 1997.

BROWN, M.L., KROS, J. F., *Data Mining and the impact of missing data*, Industrial Management and Data Systems, 108, 2003.

CHAPMAN, Pete; Clinton, Julian; Kerber, Randy; Khabaza, Thomas; Reinartz, Thomas; Shearer, Colin; Wirth, Rüdiger; "*CRISP-DM 1.0 – Step-by-Step data mining guide*"; CRISP-DM Consortium, 2000

COOKE, Simon; "*Database Marketing: strategy or tactical tool?*" Marketing Intelligence & Planning, Vol 12, No 6, 2000.

CROSS, K. F.; LYNCH, R. L. *Managing the corporate warriors*. Quality Progress, Vol. 23, No. 4, apr. 1990.

DE TONI, A.; TONCHIA, S. *Performance measurements systems, models, characteristics and measures*. International Journal of Operations & Production Management, Vol. 21, No. 1-2, 2001.

FAYYAD, U, Piatetsky-Shapiro, G.; P. Smyth; Uthurusamy, R.; "*Advances in Knowledge Discovery & Data Mining*", Cambridge, MA (The AAAI Press/The MIT Press), 1996.

FEELDERS, A.; Daniels, H.; Holsheimer, M. *Methodological and practical aspects of data mining*. Information & Management, Vol. 37, 2000.

GHALAYINI, A. M. & Noble, J. S. *The changing basis of performance measurement*. International Journal of Operations & Production Management. V16, No 8, dez/1996.

HUGHES, Arthur M.;"*The Complete Database Marketer*"; Chicago; Probus Publishing Co, 1995.

KAPLAN, R. S.; NORTON, D. *Why does Business Need a Balanced Scorecard*. Journal of Cost Management. Vol. 11, No. 3, 1997.

KENNERLEY, M.; NEELY, A. *Measuring performance in a changing business environment*. International Journal of Operations & Production Management , Vol. 23, No. 2, 2003.

LAROSE, Daniel T. *Data Mining Methods and Models*, a John Wiley & Sons, inc; 2006.

LO, Victor S.; "*The True Lift Model - A Novel Data Mining Approach to Response Modeling in Database Marketing*"; SIGKDD Explorations; Vol 2 No2; 2002.

MATTOZO, T.C. 2007. *Análise de Desempenho de Vendas em Telecomunicações Utilizando Técnicas de Mineração de Dados* (Dissertação de Mestrado). Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal do Rio Grande do Norte, Natal, 2007.

MICKEY, Ruth M. Dunn, O. & CLARCK, V., *Applied statistics: analysis of variance and regression*, John Wiley & Sons, 2005.

NEELY, A.; Adams, C.; Crowe, P. *The performance prism in practice*. Measuring Business Excellence, Vol. 5, No. 2, 2001.

NETER, J.; WASSERMAN, W. *Applied Linear Statistical Models*. Richard D. Irwin, Inc, Illinois, 1974.

NORUSIS, M.. *SPSS 13.0 Guide to Data Analysis*. Upper Saddle-River, N.J.: Prentice Hall, Inc.. 2004.

NORREKLIT, H.. *The balance on the balanced scorecard - a critical analysis of some of its assumptions*. Management Accounting Research, London, Vol.11, No.1, 2000.

O'GUIN, C. Michael; Ogilvie, Timothy; "*The Science, Not Art, of Business Intelligence*", Competitive Intelligence Review, vol. 12(4), 2001.

SPSS (2007). *Statistical Package for Social Sciences*. SPSS v. 13.0. URL: http://www.spss.com. Acesso em agosto, 2007.

UTHURUSAMY R.; FAYYAD, U.; "*Evolving data mining into solutions for insights*". Communications of the ACM 45 (8); 2002.

WELGE, Michael E.; SHAW, Michael J.; Subramaniam, Chandrasekar; Tan, Gek Woo " *Knowledge management and data mining for marketing*"; Decision Support Systems, Vol. 31 No 1, 2001.

WIRTH, Ruediger; "*CRISP-DM Position Statement*", 6th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, USA, 2000.

ZHOU, Z. H. *Three perspectives of data mining*. Artificial Intelligence jounal 143(1), 2003.

ZWICK, Detlev, Nikhilesh Dholakia "*Whose Identity Is It Anyway*? Consumer Representation in the Age of Database Marketing"; Journal of Macromarketing, Vol. 24, No. 1, 2004.