



Universidade Federal de Uberlândia
Faculdade de Matemática

Bacharelado em Estatística

UM MODELO LOGÍSTICO BAYESIANO PARA ESTIMAR A
PROBABILIDADE DE DESLIGAMENTO DE VENDEDORES DO
SETOR DE TELEMARKETING DE UMA EMPRESA

Victor Moreira Gonçalves

Uberlândia-MG

2017

Victor Moreira Gonçalves

**UM MODELO LOGÍSTICO BAYESIANO PARA ESTIMAR A
PROBABILIDADE DE DESLIGAMENTO DE VENDEDORES DO
SETOR DE TELEMARKETING DE UMA EMPRESA**

Trabalho de Conclusão de Curso apresentado à Faculdade de Matemática da Universidade Federal de Uberlândia como requisito para a obtenção do grau de bacharel em Estatística.

Orientador: Professor Dr. José
Waldemar da Silva.

Uberlândia-MG

2017



Universidade Federal de Uberlândia
Faculdade de Matemática

Coordenação do Curso de Bacharelado em Estatística

A banca examinadora, conforme abaixo assinado, certifica a adequação deste trabalho de conclusão de curso para obtenção do grau de Bacharel em Estatística.

Uberlândia, ____ de _____ de 20____

BANCA EXAMINADORA

Professor Dr. José Waldemar da Silva

Professor Dr. Ednaldo Carvalho Guimarães

Professor Dr. Janser Moura Pereira

Uberlândia-MG

2017

Agradecimentos

Primeiramente, gostaria de agradecer a Deus pela oportunidade de estar passando por este momento tão esperado, um momento em que nos tornamos de fato profissionais reconhecidos, agradecer aos meus pais Edmur e Lucienne, minha irmã Paula e meu irmão Guilherme, que sem o constante apoio e estrutura que me proporcionaram, as dificuldades por mim enfrentadas nestes anos de graduação seriam bem mais complicadas.

Aos meus professores, é uma enorme satisfação dizer que pude passar incríveis cinco anos de enorme aprendizado e crescimento, tanto profissional quanto como pessoal. Especificamente a meu orientador, professor Dr. José Waldemar, que foi paciente e trabalhou bastante para desenvolvermos este trabalho.

Aos meus amigos, só tenho a agradecer, foram vários momentos de diversão e ao mesmo tempo de dificuldades, mas que devido a nossa união e companheirismo, superamos todos os problemas enfrentados. Vocês estarão sempre em minhas lembranças onde quer que estejam.

Aos profissionais e amigos da Martins Atacadista e Distribuidor SA, sem vocês, com toda certeza eu não seria um terço do profissional que me tornei. Foram momentos essenciais em minha vida até o presente momento.

À Dra. Daniele do Prado, minha madrinha, que me indicou para o curso de estatística, saiba que você foi fundamental para que eu tivesse esta experiência dos últimos anos, seu apoio e entusiasmo sempre foram um diferencial.

À todos que me auxiliaram de alguma forma, sou e serei eternamente grato.

Resumo

No setor de telemarketing de uma empresa do ramo atacadista, constantemente lida-se com altos índices de rotatividade dos vendedores, acarretando em altos custos com rescisões, treinamentos, processos seletivos, entre outros. Considerando que os trabalhadores geralmente retornam o investimento neles realizados após um ano trabalhado, o problema é tratado em termos de probabilidade, no qual a ideia é prever probabilisticamente quando um vendedor está propenso a se desligar. Para tanto, fez-se o uso de um modelo logístico, tratando como evento resposta quando os operadores de telemarketing são desligados com menos de um ano trabalhado. Foram consideradas 19 covariáveis na análise. A estimativa dos parâmetros do modelo foi realizada via inferência bayesiana. Nesta metodologia o conhecimento prévio é atualizado utilizando a informação oriunda dos dados, alocada na verossimilhança. Admita-se que o conhecimento prévio sobre os parâmetros seja descrito por distribuições de probabilidades. A informação prévia dos dados é traduzida em uma distribuição *a posteriori* por meio do teorema de Bayes. Como a distribuição *a posteriori*, em geral, tem forma complexa, obtém-se amostras das distribuições condicionais completas para cada parâmetro iterativamente e recursivamente utilizando algoritmos que faz o uso do método de Monte Carlo via cadeia de Markov. Quando as cadeias atingirem convergência, as amostras das condicionais completas representam satisfatoriamente e empiricamente as distribuições *a posteriori*. O estudo de convergência foi feito utilizando os critérios de Geweke (GEWEKE, 1992) e critério de Raftery e Lewis (RAFTERY & LEWIS, 1992). Para selecionar as variáveis importantes na ocorrência do evento de interesse, baseou-se no teste de hipóteses genuinamente bayesiano (FBST) e em seu e-valor. O procedimento para seleção de variáveis foi semelhante ao *backward* na inferência clássica. A partir do modelo ajustado é possível prever a probabilidade de desligamento de um vendedor, sendo necessário para isto, apenas conhecer o seu perfil quanto a idade, faturamento médio e tempo em que permanece em chamada, haja vista que estas foram as variáveis selecionadas para compor o modelo final. A assertividade do modelo foi de 73% e portanto, com base neste critério, o ajuste foi considerado adequado.

Palavras-chave: *Inferência Bayesiana, Seleção de variáveis, Razão de chances, Resposta binária.*

Abstract

In the telemarketing sector of a wholesaler company, it is constantly dealing with high turnover rates, resulting in high costs with remediation, training, selective processes, among others. Considering that employees usually return, after a year worked, the investment made for them, the problem is treated in terms of probability, in which the idea is to predict probabilistically when a seller is prone to departure. For that, a logistic model was used, treating as an answer event when the telemarketers are dismissed with less than a year worked. 19 covariates were considered in the analysis. The model parameters were estimated using Bayesian inference. In this methodology the previous knowledge is updated using the information derived from the data, allocated in the likelihood. It is assumed that prior knowledge about the parameters is described by probability distributions. The previous information of the data is translated into an a posteriori distribution through the Bayes' theorem. As *the posterior* distribution, in general, has a complex form, samples of the complete conditional distributions are obtained for each parameter iteratively and recursively using algorithms that makes use of the Monte Carlo method by Markov chain, when the algorithm reaches convergence, the Conditional samples represent satisfactorily and empirically *the posteriori* distributions. The conversion study can be done using the Geweke criterion (Geweke, 1992) and Raftery and Lewis criterion (RAFTERY & LEWIS, 1992). To select the important variables in the occurrence of the event of interest, it was based on the genuinely Bayesian hypothesis test (FBST) and its e-value. The procedure for selecting variables was similar to the backward in classical inference. From the fitted model it is possible to predict the probability of a seller being disabled, being necessary for this, only to know their profile based on age, average billing and time in which they remain in call, since these were the selected variables to compose the final model. The assertiveness of the model was 73% and therefore, based on this criterion, the adjustment was considered adequate.

Keywords: *Bayesian Inference, Variable Selection, Odds Ratio, Binary Response.*

Sumário

1	Introdução	1
2	Fundamentação teórica	3
3	Metodologia	5
4	Resultados e Discussões	14
5	conclusões	40
6	Referências	41
7	Apêndice	44

Lista de Figuras

1	Ilustração do valor-e para aplicação do FBST	12
2	Boxplot para as variáveis quantitativas, x_1 a x_{12}	15
3	Gráficos para convergência das cadeias para os parâmetros de β_0 a β_{19}	17
4	Densidade das cadeias para os parâmetros de β_0 a β_{19}	19
5	Plot entre as variáveis x_3 a x_8 para confirmação da existência de multicolinearidade.	37
6	Plot dos valores preditos contra os observados.	39

Lista de Tabelas

1	Codificação das variáveis.	6
2	Tabela de contingência para variáveis qualitativas.	16
3	Estatísticas dos testes de convergências das cadeias para cada parâmetro.	21
4	Estatísticas das cadeias dos parâmetros para o primeiro passo da seleção de variáveis	22
5	Estatísticas das cadeias dos parâmetros para o segundo passo da seleção de variáveis	24
6	Estatísticas das cadeias dos parâmetros para o terceiro passo da seleção de variáveis	25
7	Estatísticas das cadeias dos parâmetros para o quarto passo da seleção de variáveis	26
8	Estatísticas das cadeias dos parâmetros para o quinto passo da seleção de variáveis	27
9	Estatísticas das cadeias dos parâmetros para o sexto passo da seleção de variáveis	28
10	Estatísticas das cadeias dos parâmetros para o sétimo passo da seleção de variáveis	29
11	Estatísticas das cadeias dos parâmetros para o oitavo passo da seleção de variáveis	30
12	Estatísticas das cadeias dos parâmetros para o nono passo da seleção de variáveis	31
13	Estatísticas das cadeias dos parâmetros para o décimo passo da seleção de variáveis	32
14	Estatísticas das cadeias dos parâmetros para o décimo primeiro passo da seleção de variáveis	33
15	Estatísticas das cadeias dos parâmetros para o décimo segundo passo da seleção de variáveis	34
16	Estatísticas das cadeias dos parâmetros para o décimo terceiro passo da seleção de variáveis	34
17	Estatísticas das cadeias dos parâmetros para o décimo quarto passo da seleção de variáveis	35
18	Estatísticas das cadeias dos parâmetros para o décimo quinto passo da seleção de variáveis	35
19	Estatísticas das cadeias dos parâmetros para o décimo sexto passo da seleção de variáveis	36
20	Estatísticas das cadeias dos parâmetros para o novo modelo sem x_8 (décimo sétimo passo)	37
21	Estatísticas das cadeias dos parâmetros para o modelo final com x_2 , x_3 e x_{10} (décimo oitavo passo)	38

1 Introdução

Muitas empresas em todo o país e de diversas áreas da economia apresentam altas taxas de *turnover* em alguns de seus setores, principalmente aqueles que demandam grande quantidade de mão de obra e apresentam remuneração relativamente baixa, geralmente próxima a um salário mínimo.

Turnover é um dos principais indicadores utilizado por líderes e pelas áreas de Desenvolvimento Humano & Organizacional de grandes companhias, que se trata do movimento de ingresso (Admissão) e evasão (Demissão) dos colaboradores de uma empresa, que pode acontecer por meio de iniciativa do colaborador ou da empresa segundo Pomi (2005).

O turnover pode vir a ocorrer por diversos fatores como ofertas mais atraentes de empresas concorrentes, instabilidade econômica, ambiente e imagem organizacional, insatisfação quanto a política salarial da empresa, poucos benefícios concedidos, condições de trabalho, política de recrutamento inadequada e outros mais conforme Rosário (2006).

É válido ressaltar, que taxas muito baixas de rotatividade podem contribuir de forma negativa também, segundo Mobley (1992, p. 30) a rotatividade afeta de certa forma o clima organizacional no relacionamento das pessoas, de forma que baixos índices não são saudáveis e podem acarretar em problemas para uma organização, como a não revitalização ou oxigenização da mão de obra, impactando na obtenção de novos talentos e novas visões do mercado para a empresa. Assim um índice saudável deve ser mantido, o qual pode variar entre empresas, negócios e áreas.

Geralmente, o que afeta a taxa de turnover é a taxa de desligamento, pois a maioria das admissões ocorrem para repor a mão de obra que está em evasão, caso a empresa não esteja em um ritmo acelerado de crescimento e expansão de “*head count*”.

Identificar colaboradores propensos a se desligarem pode não ser algo simples e fácil, por isso a atribuição de modelos preditivos torna-se uma opção viável para relacionar covariáveis que traduzem melhor o fator de desligamento.

Se a resposta for tratada como variável binária, de modo que o valor “1” represente o desligamento e “0” represente a permanência do colaborador, podemos tratar esta resposta em termos de probabilidade, realizando o ajuste de um modelo logístico. Alguns autores como Hosmer e Lemeshow (1998), Ebrahim (2001) e Guimarães (2001) utilizam da metodologia de regressão logística para modelagem estatística quando se tem uma variável categórica ou binária como resposta.

O ajuste de modelo sob enfoque bayesiano tem sido cada vez mais utilizado. Na abordagem

bayesiana a informação dos parâmetros de interesse são atualizados por meio da informação oriunda do experimento ou dados amostrais. O conhecimento prévio alocado em uma distribuição de probabilidade denominada distribuição *a priori*, a partir da qual expressa-se a incerteza, anterior à coleta dos dados, sobre o parâmetro ou sobre a quantidade de interesse. Haja vista a escolha de uma distribuição de probabilidade para o parâmetro, sob esta ótica, ele é tratado como uma variável aleatória.

A distribuição *a priori* será mais informativa quanto menor for a variância *a priori*. Além da variação, outras características devem ser observadas no ato da escolha da distribuição *a priori*, como espaço paramétrico, simetria e maior ou menor densidade em determinadas regiões do espaço paramétrico.

A informação prévia pode vir de um estudo anterior, de algum especialista ou até mesmo do senso comum.

Combinando esta distribuição *a priori* com a função de verossimilhança utilizando o teorema de Bayes determina-se a distribuição *a posteriori*.

A abordagem bayesiana é mais vantajosa no sentido de que não é dependente do tamanho da amostra e permite determinar distribuições para os próprios parâmetros e não para seus estimadores o que ocorre na inferência paramétrica clássica. Além disso, a inferência bayesiana proporciona uma interpretação probabilística para os intervalos de credibilidade, nos quais está incorporada a informação contextual específica *a priori*, enquanto que os intervalos de confiança (inferência clássica) são baseado nos dados apenas.

Toda inferência bayesiana é realizada a partir da distribuição *a posteriori* e desta forma, é necessário que tal distribuição seja determinada. Na maioria das situações a distribuição *a posteriori* é multivariada e o cálculo analítico de medidas resumo, como valor-esperado ou vetor de valores esperados, assim como a matriz de variâncias e covariâncias, não são facilmente determinados. Este problema é resolvido utilizando métodos iterativos como o método de simulação de Monte Carlo via cadeia de Markov, que na língua inglesa é escrito como *Markov Chain Monte Carlo* (MCMC) para obter amostras da(s) distribuição(ões) *a posteriori*.

Devido ao número de integrais e a complexidade de resolução, é viável que se calcule a distribuição *a posteriori* pelo método de simulação de Monte Carlo Cadeias de Markov (MCMC), cuja ideia básica é simular uma cadeia de Markov até que a distribuição de equilíbrio ou de interesse seja atingida. Esta simulação ocorre pelo método iterativo de amostragem de Gibbs (GELFAND et al., 1990).

Além de se obter um modelo adequado, é importante que a influência de cada variável, ou seja, as estimativas dos parâmetros encontradas sejam avaliadas pelo quão influentes elas

são na ocorrência do evento de interesse. Para tanto, por não se tratar de um modelo linear, tal influência pode ser verificada por meio do cálculo do “*odds ratio*”, ou razão de chances, que possibilita, para modelos logísticos, tratar a mudança no evento de interesse ocorrer em termos de proporções. A importância, ou relevância da razão de chances encontrada para cada parâmetro fica a critério dos analistas da área estudada.

Estima-se que um vendedor do setor de telemarketing de uma determinada empresa do ramo atacadista, retorna todo o investimento nele feito, somente um ano após sua admissão, ou seja, é quando o mesmo começa a gerar lucro, isso quando conseguem ter constância em vendas relativamente boas. Assim, é interessante determinar qual(is) dos colaboradores tem maior probabilidade de permanecer menos que um ano para estabelecer ações corretivas, pois o turnover com menos de doze meses de trabalho é sinônimo de prejuízo.

O ajuste de um bom modelo permite que a empresa adote políticas para retenção de colaboradores personalizadas para determinados públicos, forneça incentivos diferentes, se prepare economicamente para possíveis rescisões, sendo possível até identificar se há padrão no perfil daqueles que permanecem mais que um ano na empresa.

Atualmente o custo com a rescisão destes vendedores, está acima dos dez mil reais por mês, considerando os custos com processos seletivos, treinamentos e custos dos cálculos rescisórios. Se considerarmos a ocorrência de custos com processos judiciais, que eventualmente ocorrem, os valores gerados pela rotatividade se elevam ainda mais.

Tratamento dos dados que se refere a modelagem, nunca havia sido realizado nesta empresa, de modo que as tomadas de decisões eram baseadas na percepção dos líderes da organização e tratamento básico dos dados, como construção de indicadores e painéis com dados gráficos consolidados.

Com isso, o objetivo geral deste trabalho é encontrar o melhor ajuste de modelo de regressão logística bayesiana, que preveja de forma parcimoniosa, ou seja, obtenha bons resultados com o mínimo de erro e também com o modelo menos complexo possível, a probabilidade de um determinado colaborador se desligar antes de completar um ano de trabalho (prejuízo).

2 Fundamentação teórica

A ideia sobre marketing por telefone em grande escala, começou a ser implementada em 1970 pela Ford Motor Company, quando a empresa contratou cerca de quinze mil mulheres para buscar novos clientes via telefone, segundo Theodore Levitt (LEVITT, 1985.apud DANTAS,1997,p.55).

No Brasil, este tipo de atividade foi difundida entre meados da década de 80 e início da década de 90, com fim do sistema Telebrás, conforme ABT (ABT,2011).

De acordo com a VI PESQUISA ABT, o telemarketing se tornou o meio de contato mais utilizado para firmar e manter relacionamento com clientes e que gera diversas vagas de emprego para profissionais sem experiência no mercado de trabalho.

De acordo com Rocha e Aguirella (2016, p.373) o ramo de telemarketing produz, constantemente, um alto número de vagas de empregos em todo o Brasil, mas que vem sempre acompanhado com altos índices de rotatividade, fazendo com que esta área de atuação seja foco de estudos frequentemente.

É comum observar que o mercado de telemarketing investe bastante em fatores incomuns quando se trata de perfil de contratação, como iniciantes na corrida profissional, pessoas com mais idade e que geralmente não conseguem emprego fácil, opção sexual e afastados do mercado de trabalho, conforme Ziliotto e Oliveira (2014, p.173).

Quando metas são difíceis de se alcançar ou injustas, podem provocar insatisfação e em algumas pessoas desenvolver a sensação de fracasso, conforme Silma Venco (VENCO, 2009).

Remuneração coerente é um fator que pode evitar que um funcionário fique desmotivado, segundo Herzberg (HERZBERG et al, 1959.apud ROBBINS, 2002), junto a este, Braga (BRAGA, 2009) discute que baixa remuneração aliado a condições ruins de trabalho, são fatores que acarretam em rotatividade.

No ponto de vista de Chiavenato (2009), a rotatividade provoca custos com rescisões, reposição da mão de obra perdida, perda de qualidade devido a saída do profissional experiente e outros efeitos colaterais complementares a médio e longo prazo.

No conceito de Assis (2005), a rotatividade se tornou um paradigma para gestores em geral, devido ao impacto financeiro causado, sendo tal impacto proporcional ao nível hierárquico. Neste caso, apesar do nível hierárquico ser baixo, a quantidade de desligamentos faz com que o gasto seja elevado.

Dada a importância do problema de rotatividade em empresas de telemarketing, o trabalho tem como propósito estudar e identificar variáveis que possam estar associadas ao desligamento de colaborador por meio da regressão logística via inferência bayesiana.

JUBILATO (2011), afirma que baixos índices de rotatividade não é sadio para as organizações, pois retrata uma organização estagnada, uma vez que não renova ou “oxigena” sua força de trabalho, não permitindo que pessoas novas, com novas metodologias e conhecimentos possam contribuir para evolução do negócio.

Patewal e Keimar (2009) afirmam que algumas técnicas estatísticas são utilizadas com

frequência para previsão e classificação do evento estudado, como regressão logística e análise de discriminante.

Problemas de classificação de variáveis em duas categorias são vistos em diversas áreas com muitas aplicações. Vallejos e McKinnon (2013) fazem a classificação de registros sísmicos, utilizando da técnica de regressão logística para tratar a ocorrência do evento de interesse em termos de probabilidade.

Na percepção de Druseitil e Ohno-Machado (2002) modelos de regressão logística está entre os mais populares, com presença em cerca de 72% das publicações na base MEDLINE.

McInturff *et al.*(2004), discutem vantagens da abordagem bayesiana sobre a abordagem clássica comumente usada no ajuste de modelos para dados binários, como a atribuição de informações *a priori* que sejam relevantes, além da possibilidade de inferir sobre amostras pequenas.

A abordagem clássica utiliza da informação amostral para estimar e testar hipóteses sobre parâmetros, ao passo que no âmbito bayesiano toda a informação amostral é utilizada, além de permitir julgamento pessoal (informação prévia) para escolha de alguma distribuição de probabilidade *a priori* para realizar o cálculo das estimativas conforme Santana *et al.*(2002) e Silva e Suárez (2000).

Conforme Brasil(1991), a análise bayesiana quando utiliza prioris distintas, pode chegar a respostas distintas, por isso, esse problema pode ser resolvido usando de prioris não informativas para obter “Objetividade”.

A tratativa bayesiana pode ser vista como extensão da clássica, uma vez que os resultados podem ser iguais, mas que são interpretados de forma diferente.

3 Metodologia

A base de dados consiste em uma série histórica de 1540 indivíduos e 19 variáveis explicativas, sendo 12 quantitativas e 7 qualitativas, coletadas no período de maio de 2014 a março de 2017.

A variável de interesse consiste em uma variável binária, cuja resposta é “um” quando o colaborador pede demissão ou é demitido com menos de um ano de trabalho e “zero” caso contrário. Estes vendedores foram avaliados mensalmente por diversas variáveis de negócio que podem influenciar no desligamento dos mesmos, sendo elas consideradas como possíveis covariáveis para o modelo a ser ajustado.

As 12 covariáveis quantitativas que inclui variáveis de negócio e caracterização dos indi-

víduos, bem como as variáveis qualitativas a serem avaliadas no modelo são apresentadas na Tabela 1.

Tabela 1: Codificação das variáveis.

Descrição	Notação
Tempo de casa (meses)	x_1
Idade (anos)	x_2
Faturamento médio	x_3
Quantidade de pedidos médio	x_4
SKU médio	x_5
Índice do vendedor	x_6
Prêmio médio por economia	x_7
Prêmio médio	x_8
Quantidade média de chamadas efetivas	x_9
Tempo médio em chamadas (segundos)	x_{10}
Média da nota de qualidade	x_{11}
Taxa média de conversão	x_{12}
Casado (Estado Civil)	x_{13}
Solteiro (Estado civil)	x_{14}
Fundamental incompleto (Escolaridade)	x_{15}
Fundamental completo (Escolaridade)	x_{16}
Ensino Médio Completo (Escolaridade)	x_{17}
Gênero	x_{18}
Dependente econômico	x_{19}

A seguir tem-se a descrição mais detalhada das variáveis apresentadas na Tabela 1.

O faturamento médio mensal (x_3) dado o tempo em que o colaborador está na empresa. Tal indicador é considerado importante para a análise, uma vez que os valores arrecadados em venda é o que movimenta a empresa e o ganho dos vendedores.

A quantidade de pedidos realizados pelo vendedor em média(x_4), ou seja, média mensal da quantidades realizadas.

A quantidade média de SKU (x_5) é a média mensal da média do número de itens distintos dentro de cada pedido dentro de um mês específico.

O índice do vendedor (x_6) é o percentual relativo ao cumprimento das metas. A cada

mês, os vendedores são confrontados com uma meta mensal de faturamento, caso tal objetivo seja atingido é contabilizado uma unidade neste indicador. Dessa Maneira, para o calculo, foi somada uma unidade toda vez que o vendedor atingiu a meta desde que iniciou na empresa dividido pela quantidade de meses do período.

O prêmio por economia de descontos médio (x_7) é a média do prêmio recebido pelos vendedores, quando estes economizam margem de lucro nos pedidos não cedendo descontos permitidos aos clientes.

Os vendedores recebem um percentual relativo as vendas realizadas dentro de cada mês, sendo este percentual maior quando os mesmos batem meta, caso contrario o percentual sobre a receita realizada é menor. Deste modo, espera-se que aqueles que conseguem bons prêmios, apresentem tendência a permanecer mais tempo na empresa, assim, foi extraída a média mensal das premiações (x_8) de cada vendedor que já trabalhou ou trabalha na empresa.

Dentre todas as ligações realizadas, grande parte não é atendida pelos clientes, das que são atendidas, algumas duram pouco tempo e o vendedor não consegue ao menos ofertar produtos. Quando estas chamadas duram mais que um tempo específico, considera-se que o vendedor conseguiu contato efetivo com o cliente. Por isso, a métrica deste indicador é baseada na média mensal do número de chamadas efetivas (x_9).

Junto a quantidade de chamadas, espera-se que o tempo em que o vendedor permanece com o cliente em ligação seja relevante, pois a empresa preza o relacionamento com o cliente, ou seja, há um senso comum de que mais tempo em ligações reflita em um bom relacionamento, consequentemente vendas melhores, porém bons vendedores conseguem bons pedidos no menor tempo possível para ganhar em produtividade. Desta forma, a média(x_{10}) do tempo médio mensal por chamada é uma variável que pode inferir na permanência dos colaboradores na empresa por mais que um ano.

Por trás de toda a equipe de vendedores, existe a equipe de qualidade que sempre acompanha e faz auditoria das ligações dos vendedores para identificar se a abordagem do mesmos segue os padrões de treinamento e se o vendedor está prezando o relacionamento com o cliente. Para tanto, os avaliadores da equipe de qualidade atribuem nota aos vendedores conforme estes vão atendendo os requisitos mínimos de uma ligação. Por tanto, a métrica desta variável é composta por uma média mensal das notas de qualidade (x_{11}) de cada colaborador.

E por fim, considera-se a média das taxas de conversão (x_{12}) obtidas em cada mês desde que o vendedor ingressou na companhia. A taxa de conversão é o percentual de ligações convertidas em vendas.

Junto a essas variáveis de negócio tem-se também algumas que caracterizam os vendedores,

como sexo e presença de dependentes econômicos, sendo estas binárias em que o valor “um” caracteriza a existência da característica. As variáveis “idade” (x_2) representa o tempo de vida dos colaboradores em anos e “tempo de casa” (x_1) que trata o tempo em que o colaborador está na empresa em meses.

A variável qualitativa “Estado civil” é composta por três classificações, “Casado” (x_{13}), “Solteiro” (x_{14}) e “Divorciado”. Esta variável será tratada no ajuste do modelo por meio de duas variáveis binárias x_{13} e x_{14} , de modo que é atribuído valor “um” em x_{13} para a existência da característica “Casado” e “zero” caso seja outra caracterização, da mesma forma, em x_{14} quando tem-se o valor “um”, para a presença da característica “Solteiro” e “zero” caso contrário. Quando em ambas as variáveis ocorrer o valor “zero”, tem-se a característica “Divorciado”

A variável “Escolaridade”, composta por quatro níveis, “Ensino fundamental incompleto” (x_{15}), “Ensino fundamental completo” (x_{16}), “Ensino médio completo” (x_{16}) e “Ensino superior completo”. Da mesma forma, os níveis serão tratados no modelo como variável binária, sendo necessárias três variáveis para tanto, x_{15} , x_{16} e x_{17} , tal que em x_{15} o valor “um” sinaliza a característica “Ensino fundamental incompleto” e “zero” caso seja outra característica. Para x_{16} , quando o valor “um” é assumido, tem-se a caracterização de “Ensino fundamental completo” para o indivíduo, do contrário, “zero” traz alguma outra classificação. Em x_{17} , “um” é atribuído quando o indivíduo possui “Ensino médio completo” e “zero” caso a característica seja alguma outra. Por fim, no momento em que as três variáveis assumirem valor “zero”, significará a existência da característica “Ensino superior completo”.

Sabe-se que o “*feeling*” dos analistas de negócio responsáveis por esta área da empresa para a detecção de quais variáveis são importantes para tal modelagem pode ser suficiente para testá-las no modelo, porém recomenda-se que seja feito ao menos uma avaliação gráfica das variáveis quantitativas e contigencial das variáveis qualitativas, a fim de que seja avaliada, mesmo que seja apenas para conhecimento, a existência de diferenças entre os eventos analisados, que justifique trabalhar com determinada variável.

Tal descrição será realizado utilizando o box plot das variáveis quantitativas e uma tabela de contigência para as variáveis qualitativas para cada grupo analisado a fim de conhecer melhor as variáveis, ou seja, explorá-las inicialmente por meios de gráficos e proporções para ver a relação destas com a variável resposta.

Como o objetivo é trabalhar com o modelo logístico como base para cálculo da probabilidade do evento de interesse ocorrer, teremos que :

$$P(Y = 1|\beta_0, \beta_2, ..., \beta_K) = \frac{e^{\beta_0 + \sum_{i=1}^{19} x_i \beta_i}}{1 + e^{\beta_0 + \sum_{i=1}^{19} x_i \beta_i}} \quad (1)$$

A expressão (1) representa a probabilidade do evento desligamento ($Y = 1$) ocorrer, dado os $k + 1$ parâmetros associados ao intercepto e às k covariáveis.

Na lógica bayesiana, espera-se que exista alguma informação sobre os parâmetros β_{is} ainda que vaga. Assumindo que tal informação possa ser expressada por meio de distribuição normal, com parâmetros ou hiperparâmetros μ_0 e σ_0^2 , a distribuição *a priori* pode ser descrita conforme expressão (2)

$$P(\beta_i) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{\frac{-(\beta_i - \mu_0)^2}{2\sigma_0^2}} \quad (2)$$

em que μ_0 e σ_0^2 são os parâmetros *a priori* definidos previamente pelo pesquisador.

Supondo independência entre os parâmetros e tratando $\beta = [\beta_1, ..., \beta_k]$, a *a priori* conjunta será como apresentado na expressão (3).

$$P(\beta_1, ..., \beta_k) = P(\beta) = \prod_{i=1}^k \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{\frac{(\beta_i - \mu_0)^2}{2\sigma_0^2}} \quad (3)$$

Sendo a variável resposta de interesse dicotômica, pode-se assumir uma distribuição Bernoulli para a mesma, cuja probabilidade de sucesso é dada por um modelo logístico como em (1). A verossimilhança se dará como em (4).

$$\begin{aligned} P(Y|\beta_1, \beta_2, ..., \beta_k) &= L(\theta; y_1, ... y_n) \\ P(Y|\beta_1, \beta_2, ..., \beta_k) &= \prod_{i=1}^n f(y_i; \theta) \\ P(Y|\beta_1, \beta_2, ..., \beta_k) &= \prod_{i=1}^n p^{y_i} (1 - p)^{1-y_i} \\ P(Y|\beta_1, \beta_2, ..., \beta_k) &= p^{\sum_{i=1}^n y_i} (1 - p)^{\sum_{i=1}^n (n-y_i)} \end{aligned} \quad (4)$$

em que p representa a probabilidade de sucesso definida em (1) e n é o tamanho da amostra ou quantidade de colaboradores utilizados neste estudo.

Assim, pelo teorema de Bayes pode-se calcular a distribuição *a posteriori* conjunta utilizando (2) e (4) para obter $P(\beta_1, \beta_2, ..., \beta_k|Y)$, como pode-se observar em (5)

$$P(\beta_1, \beta_2, \dots, \beta_k | Y, X) = \frac{P(\beta_1, \beta_2, \dots, \beta_k) P(Y | \beta_1, \beta_2, \dots, \beta_k)}{P(Y, X)} \quad (5)$$

Sendo assim,

$$P(\beta_1, \dots, \beta_k | Y, X) = \frac{\prod_{i=1}^k \left(\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(\beta_i - \mu_0)^2}{2\sigma_0^2}} \right) p^{\sum_{j=1}^n y_j} (1-p)^{\sum_{j=1}^n (n-y_j)}}{\int_{\beta_0}, \dots, \int_{\beta_k} \prod_{i=1}^k \left(\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(\beta_i - \mu_0)^2}{2\sigma_0^2}} \right) p^{\sum_{j=1}^n y_j} (1-p)^{\sum_{j=1}^n (n-y_j)} d\beta_1, \dots, d\beta_k} \quad (6)$$

Tratando $\beta = [\beta_1, \beta_2, \dots, \beta_k]$ e substituindo as distribuições em questão, tem-se a expressão (7) referente à distribuição *a posteriori*:

$$P(\beta | Y, X) = \frac{\prod_{i=1}^k \left(\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(\beta_i - \mu_0)^2}{2\sigma_0^2}} \right) p^{\sum_{j=1}^n y_j} (1-p)^{\sum_{j=1}^n (n-y_j)}}{\int_{\beta} \prod_{i=1}^k \left(\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(\beta_i - \mu_0)^2}{2\sigma_0^2}} \right) p^{\sum_{i=1}^n y_i} (1-p)^{\sum_{i=1}^n (n-y_i)} d\beta} \quad (7)$$

A expressão contida no denominador em (7) seria composta por $k+1$ integrais, quantidade de parâmetros do modelo, tornando inviável os cálculos desta constante normalizadora o que afetaria a determinação das marginais. A marginal para um determinado parâmetro demandaria a integração da distribuição *a posteriori* conjunta em relação aos demais parâmetros o que pode ser impossível. Alternativamente, amostra-se das distribuições condicionais completas *a posteriori* e havendo convergência das cadeias geradas, tem-se que estas distribuições se aproximam satisfatoriamente das marginais correspondentes. O trato matemático da distribuição *a posteriori* para obtenção das distribuições condicionais completas *a posteriori* é relativamente mais fácil e além disso, existem recursos computacionais que auxiliam na obtenção destas amostras. Neste trabalho em específico será utilizando o pacote rjags que permite uma interação entre os softwares R (R Core Team, 2017) e JAGS (PLUMMER, 2011).

Para tanto, comumente utiliza-se o amostrador de Gibbs, cuja função é simular de distribuições multivariadas, muito utilizado no âmbito bayesiano para amostrar recursivamente e de forma iterativa as distribuições condicionais completas *a posteriori*, já que a distribuição conjunta é mais complicada de se calcular.

Devido a amplitude do espaço paramétrico dos parâmetros do modelo, os hiperparâmetros da distribuição normal *a priori* foram considerados como $[\mu_0; \sigma_0^2] = [0; 100000]$. Considera-se que este valor de variância defina uma distribuição pouco informativa previamente.

As cadeias geradas para cada parâmetro devem atender a condição de estacionariedade para que representem as amostras de suas marginais. Por isso há a necessidade de verificar a convergência das cadeias dos parâmetros. Existem alguns critérios de estudo de convergência como o critério de Raftery e Lewis (RAFTERY & LEWIS, 1992), cuja regra de decisão é baseada no fator de dependência, responsável pelo acréscimo multiplicativo ao número de iterações que se precisa para obter convergência, se este for maior que “cinco” conforme Lewis (1992b), não há convergência da cadeia. Outro critério amplamente utilizado é o critério de Geweke (GEWEKE, 1992), que trata de um teste de hipóteses para comparar duas frações da cadeia, ou seja, verifica se uma fração inicial da cadeia difere significativamente de uma fração final. Se a hipótese de nulidade não for rejeitada não se rejeita também a hipótese de convergência. Assim, após ajuste do modelo, verifica-se a convergência das cadeias das distribuições condicionais completas *a posteriori*, haja vista que a representatividade das marginais por meio de amostras das condicionais está condicionada à convergência.

Após verificar convergência das cadeias responsáveis por gerarem as distribuições marginais *a posteriori*, é necessário que se identifique a significância dos parâmetros da regressão, de forma a estabelecer quais variáveis serão influentes no cálculo da probabilidade final de interesse. Uma forma possível e simples consiste em eliminar variáveis que a hipótese de nulidade para os parâmetros relacionados não seja rejeitada. Em outras palavras se o intervalo de máxima densidade *a posteriori*, do inglês “*Highest Posterior Density*” (HPD) incluir o zero, tal variável poderia excluída do modelo.

O HPD é determinado a partir de uma probabilidade fixada com a condição de que valores pertencentes a este intervalo tenha densidade superior a qualquer valor não pertencente. Sua interpretação para descarte de variáveis para um modelo é similar ao intervalo de confiança na estatística clássica, de modo que se o intervalo contiver o valor “zero”, a variável pode ser eliminada, caso contrário deve ser avaliada por algum critério bayesiano mais apurado, pois, apesar do intervalo de credibilidade possibilitar a avaliação de variáveis não influentes no modelo, não é suficiente decidir quais variáveis devem fazer parte do modelo final, visto que não é possível determinar qual variável possui maior probabilidade de ser retirada do modelo em casos nos quais mais de uma inclui o valor “zero” no HPD. O HPD não permite determinar qual variável possui maior probabilidade de ser excluída do modelo, para tanto este problema pode ser resolvido com o uso do teste de significância genuinamente bayesiano, que do inglês é chamado “*Full Bayesian Significance Test*” (FBST) proposto por Pereira *et. al* (1999).

O FBST representa a evidência contrária à hipótese nula, de forma a testá-la pelo cálculo da probabilidade *a posteriori* complementar à região intervalar de máxima densidade, que é

limitante ao conjunto de valores que descrevem a hipótese de nulidade, sendo baseada no valor-e. Esse teste é dito genuinamente bayesiano pelo fato de depender unicamente da distribuição *a posteriori*. O valor-e é a probabilidade de o parâmetro assumir um valor mais extremo do que o valor postulado na hipótese nula, sendo que esta probabilidade é calculada a partir da distribuição *a posteriori*.

Intervalo HPD que tangencia a hipótese nula, é o conjunto $T(x)$ definido em (8)

$$T(x) = \{\theta \in \Theta : h(\theta|X) > h(\theta^*|X)\} \quad (8)$$

em que θ^* é a máxima densidade da distribuição *a posteriori*, que segrega a região de não rejeição da hipótese nula da região de rejeição.

Desta forma, defini-se o valor-e como em (9).

$$valor - e = 1 - P(\theta \in T(x)|X) \quad (9)$$

Assim pode-se dizer que o valor-e consiste na região *a posteriori* pertencente ao espaço paramétrico de θ , em que se encontra os pontos que possuem densidade inferior a densidade do ponto de máximo da distribuição *a posteriori* sujeita hipótese nula.

A forma de avaliar a rejeição ou não de hipóteses de nulidade é por meio do valor-e, que é similar ao caso paramétrico, no qual utiliza do valor p em testes de hipóteses. A Figura 9 esboça a interpretação do valor-e.

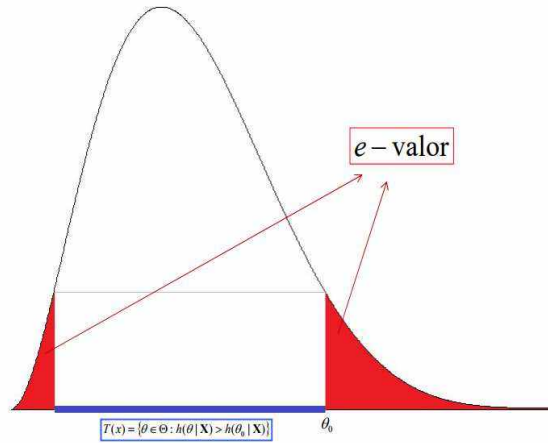


Figura 1: Ilustração do valor-e para aplicação do FBST

A seleção de variáveis por meio do teste FBST, neste trabalho, é similar ao método “Backward” no âmbito da inferência clássica, ou seja, inicia-se com o ajuste do modelo com todas as variáveis presentes, aquele parâmetro que tiver como resultado o maior valor-e entre os demais é

retirado do modelo e um novo ajuste é feito sem a respectiva variável. Assim é feito até que não reste variáveis cuja hipótese nula não seja rejeitada. Para este trabalho foi adotado o valor de 5% como probabilidade de referência para a tomada de decisão. Se o valor-e for menor que 0,05, a decisão será rejeitar a hipótese nula.

O FBST é um teste baseado em uma função analítica, ou seja quando a distribuição *a posteriori* é conhecida, neste caso, trabalha-se com uma distribuição *a posteriori* empírica, assim o valor-e será calculado também de forma empírica.

Neste caso para obter o valor-e, foi considerada a menor dentre as probabilidades de se obter um valor para o parâmetro, sejam elas a direita ou a esquerda do valor sob a hipótese nula ($\beta = 0$). Esta probabilidade duplicada define o valor-e nos casos em que as distribuições *a posteriori* são simétricas. O estudo da simetria foi realizado por meio do coeficiente de assimetria e calculado a partir da amostra para cada parâmetro, em cada um dos passos de seleção de variáveis.

Segundo Lopes (2003) uma distribuição de um conjunto de dados pode ser considerado simétrica se o coeficiente de assimetria estiver entre os valores de $-0,2$ e $0,2$. Sendo assim este critério foi adotado para o contexto deste trabalho, de forma que a hipótese de simetria não é rejeitada, caso o coeficiente de assimetria esteja no intervalo indicado.

Em modelos lineares a interpretação do efeito de um parâmetro para o modelo é bem simples, realizado de forma multiplicativa com a variável e aditiva com os demais efeitos. Tratando-se de um modelo logístico, tal interpretação passa a ser realizada pelo “*odds ratio*” (razão de chances)

De acordo com Ferreira (2012), os parâmetros de um modelo logístico podem ser interpretados separadamente, comparando a probabilidade de sucesso com a probabilidade de fracasso, por meio do odds ratio (razão de chances). Sabe-se que o modelo em questão é construído sobre o evento de sucesso, com base na seguinte função:

$$\pi(x) = \frac{e^{\beta_0 + \sum_i \beta_i x_i}}{1 + e^{\beta_0 + \sum_i \beta_i x_i}} \quad (10)$$

Como a razão de chances lida com a razão entre o sucesso e o fracasso, teremos a função odds, tal como:

$$h(x) = \frac{\pi(x)}{1 - \pi(x)} = \frac{\frac{e^{\beta_0 + \sum_i \beta_i x_i}}{1 + e^{\beta_0 + \sum_i \beta_i x_i}}}{1 - \frac{e^{\beta_0 + \sum_i \beta_i x_i}}{1 + e^{\beta_0 + \sum_i \beta_i x_i}}} = e^{\beta_0 + \sum_i \beta_i x_i} \quad (11)$$

Assim, se tomados dois valores distintos em uma ou mais unidades teremos o cálculo da razão de chances, seguindo como em (12).

$$OR = \frac{h(x_j + \Delta)}{h(x_i)} = \frac{e^{\beta_0 + \sum_{i, i \neq i'} \beta_i x_i + \beta_{i'}(x_{i'} + \Delta)}}{e^{\beta_0 + \sum_{i, i \neq i'} \beta_i x_i + \beta_{i'} x_{i'}}} = e^{\beta_{i'} \Delta} \quad (12)$$

De modo que se considerado o incremento Δ em uma das variáveis explicativas, mantendo as demais fixas, o valor de OR representará em percentual o quanto tal variável influenciará na ocorrência do evento de interesse pela expressão (13).

$$OR = e^{\beta_{i'} \Delta} \quad (13)$$

Assim, caso um parâmetro tenha estimativa positiva, a exponencial no do mesmo, ocasionará em uma razão de chances maior que “um”, de forma que a cada unidade incrementada ao valor “um”, por exemplo, uma OR igual a 2,57, para o incremento de Δ na variável explicativa, implica em 1,57 vezes mais chances de ocorrência do evento estudado. Do contrário, uma OR igual a 0,82, implica em uma redução nas chances de ocorrência do evento de 0,18 vezes.

A implementação da análise bayesiana para o ajuste do modelo logístico foi realizada via software estatístico R (R Core Team, 2017), por meio do pacote Rjags (PLUMMER, 2013) no qual contém os algoritmos citados já implementados (disponível em anexo).

4 Resultados e Discussões

Após adequação da base de dados para entrada no software R, executou-se uma avaliação prévia das variáveis quantitativas, a fim de identificar graficamente se há alguma provável diferença entre os grupos presentes na variável resposta, onde 0 é o grupo dos desligados com mais de um ano de serviço ou ativos com mais de um ano trabalhado e o grupo 1 que representa aqueles que foram demitidos com menos de um ano trabalhado ou que estão ativos, mas ainda não completaram um ano de empresa.

Para tanto, realizou-se um boxplot para cada covariável contra a variável de interesse, para conhecimento das mesmas, conforme Figura 2.

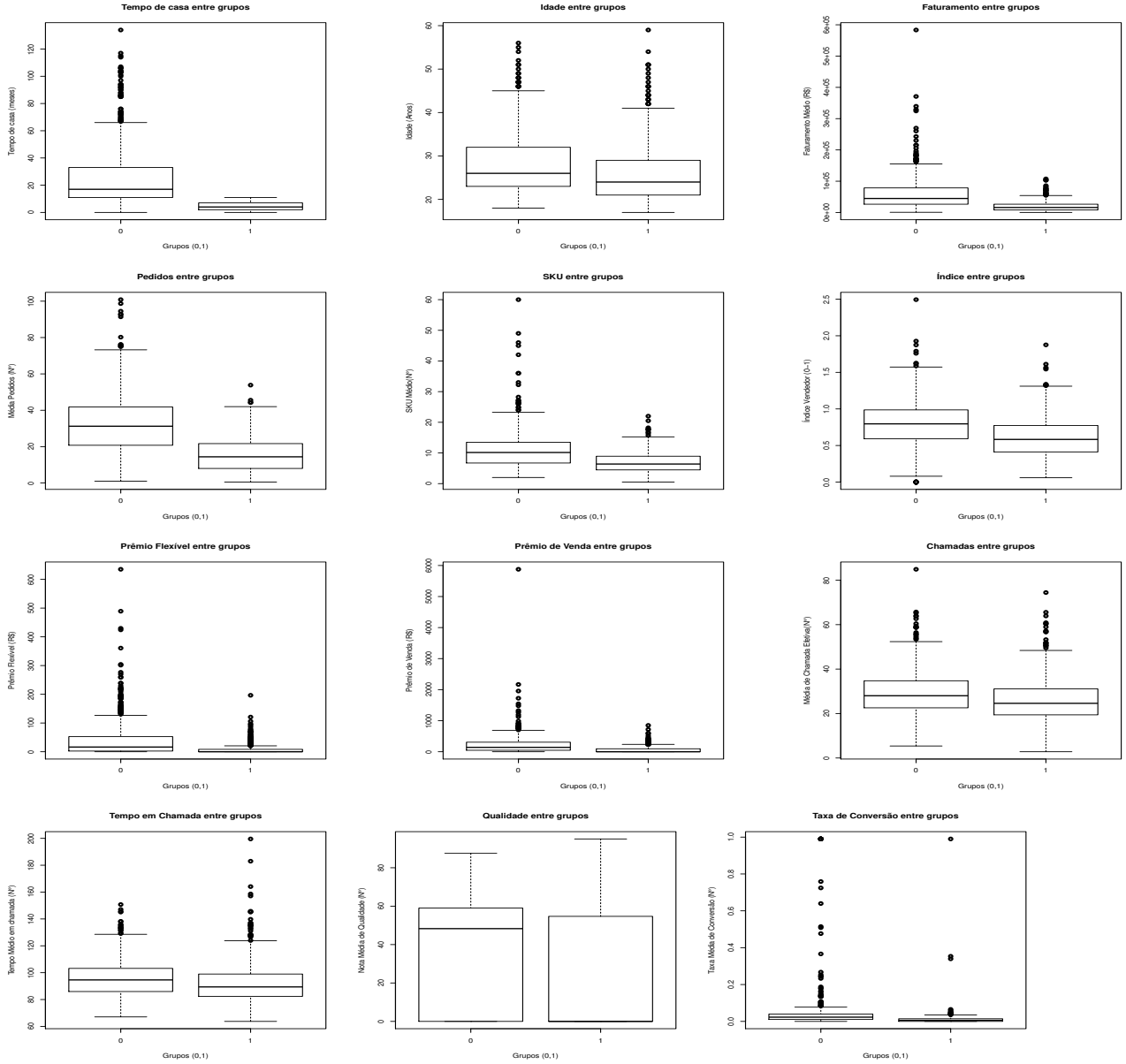


Figura 2: Boxplot para as variáveis quantitativas, x_1 a x_{12} .

Pelos gráficos plotados na Figura 2, nota-se a presença de alguns *outliers* nos grupos, principalmente nos grupos que estão ou permaneceram mais que um ano na empresa, isso se deve pela presença de bons vendedores que permaneceram ou estão a um longo período de tempo na companhia, com bons indicadores, como altas receitas, recebendo bons prêmios e assim por diante. Logo, tais *outliers* não podem ser retirados da análise, pois refletem o comportamento de vendedores de alta performance.

Assim nota-se (Figura 2) uma maior diferença entre os grupos em questão nas variáveis “Tempo de Casa”, “Faturamento”, “Média de Pedidos”, “índice do Vendedor”, “Prêmio Flexível” e “Prêmio de venda”. Com isso, há um indicativo de que estas variáveis possam ser significativas para o modelo em questão, enquanto que as demais “Idade”, “SKU”, “Chamadas efetivas”,

“Tempo médio em chamada”, “Qualidade” e “Taxa de conversão” mostram um comportamento menos divergente, mas que não impede que as mesmas sejam importantes para o modelo ajustado, ou seja, serão todas consideradas no modelo inicial.

Para as variáveis qualitativas obteve-se uma tabela contingência conforme a Tabela 2.

Tabela 2: Tabela de contingência para variáveis qualitativas.

Parâmetros	Grupos avaliados	
	Grupo 0	Grupo 1
Casados	25,03%	18,73%
Solteiros	70,82%	78,67%
Divorciados	4,15%	2,60%
Ensino Fundamental Incompleto	1,04%	1,69%
Ensino Fundamental Completo	7,26%	5,20%
Ensino Médio Completo	86,12%	88,43%
Superior Completo	5,58%	4,68%
Gênero Feminino	81,19%	65,80%
Dependente Econômico	53,44%	41,22%

Pela Tabela 2, podemos perceber que para quase todas as variáveis em questão, os grupos 0 e 1 apresentam proporções semelhantes das características em suas populações, ou seja provavelmente não implicará em mudanças na permanência ou não de colaboradores na companhia.

A variável que apresenta maiores chances de apresentar significância para o modelo a ser ajustado, poderá ser “Gênero”, pois há uma diferença relativamente notável na população de pessoas do gênero feminino entre os grupos, de modo a se pensar que o fato de um vendedor ser mulher, pode implicar em maior tempo de permanência na empresa, enquanto que homens podem ter maior facilidade em se desligarem ou serem desligados.

Para tanto, foi testado o modelo com todas as variáveis em questão e modelos sem variáveis menos influentes para identificar o melhor ajuste. Para tal análise, foi executada a rotina que junto ao pacote Rjags, permite a obtenção da estimativa dos parâmetros para a regressão.

Para ajuste dos parâmetros do modelo foi considerado um descarte (burn in) de dez mil iterações a fim de eliminar o efeito dos chutes iniciais iguais a zero, neste caso.

E por fim, foram executadas cento e vinte mil iterações com saltos de dez iterações (“*thinning*”) para garantir a convergência e estacionariedade das cadeias para as distribuições marginais ainda que empiricamente.

Inicialmente realizou-se uma análise por meio dos gráficos das séries das cadeias dos parâmetros *a posteriori*, a fim de explorar informalmente a estacionariedade das mesmas, conforme Figura 3.

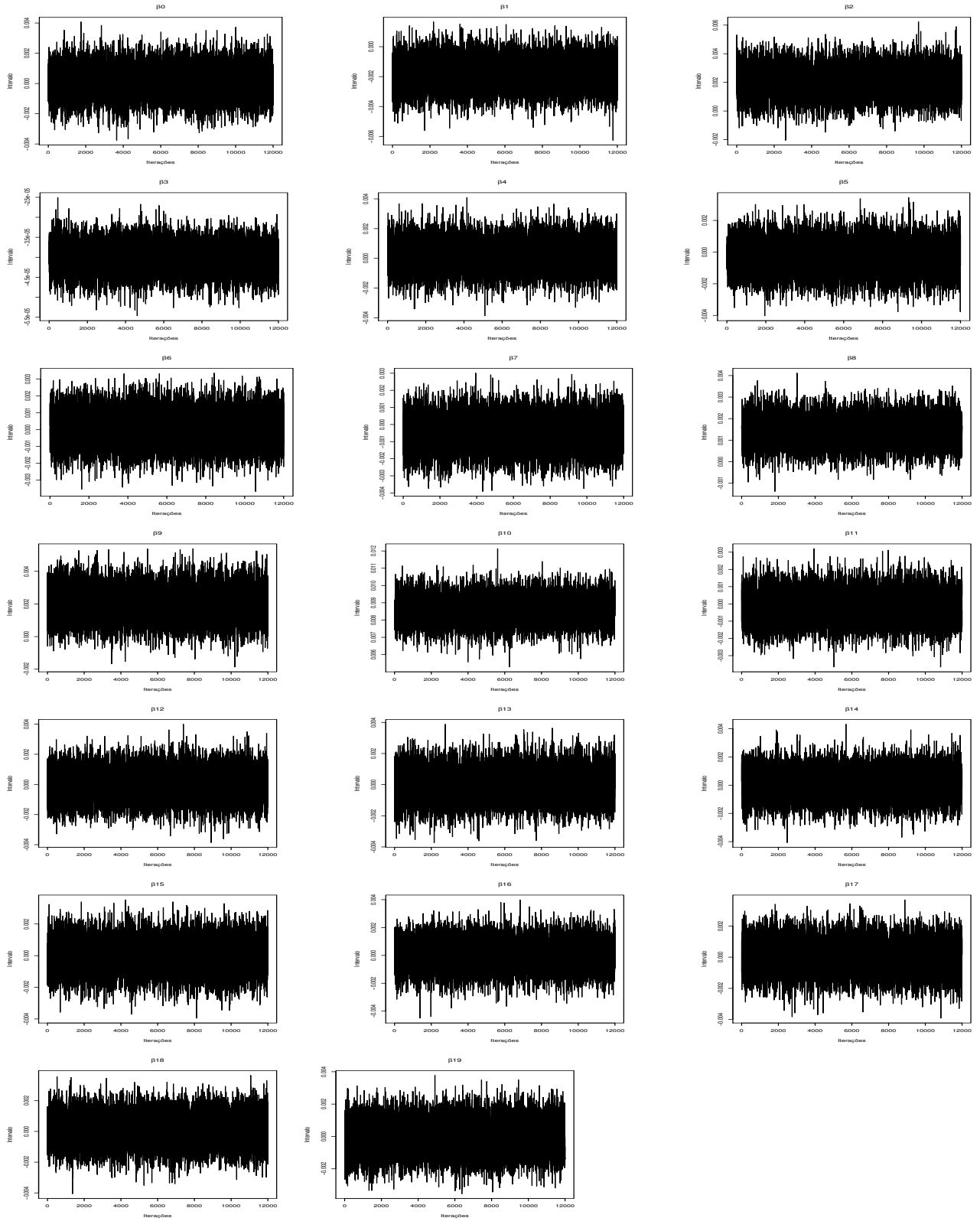


Figura 3: Gráficos para convergência das cadeias para os parâmetros de β_0 a β_{19} .

Nota-se que em todas as cadeias não é observado o efeito de tendência, ou sazonalidade, ou outro padrão que descaracterize não aleatoriedade nas séries. Logo as mesmas aparentam estar estacionárias para todos os parâmetros, ou seja, são indicativos de convergência.

Junto ao gráfico das séries das cadeias geradas pelo algoritmo MCMC, observa-se as densidades destas, conforme a Figura 4.

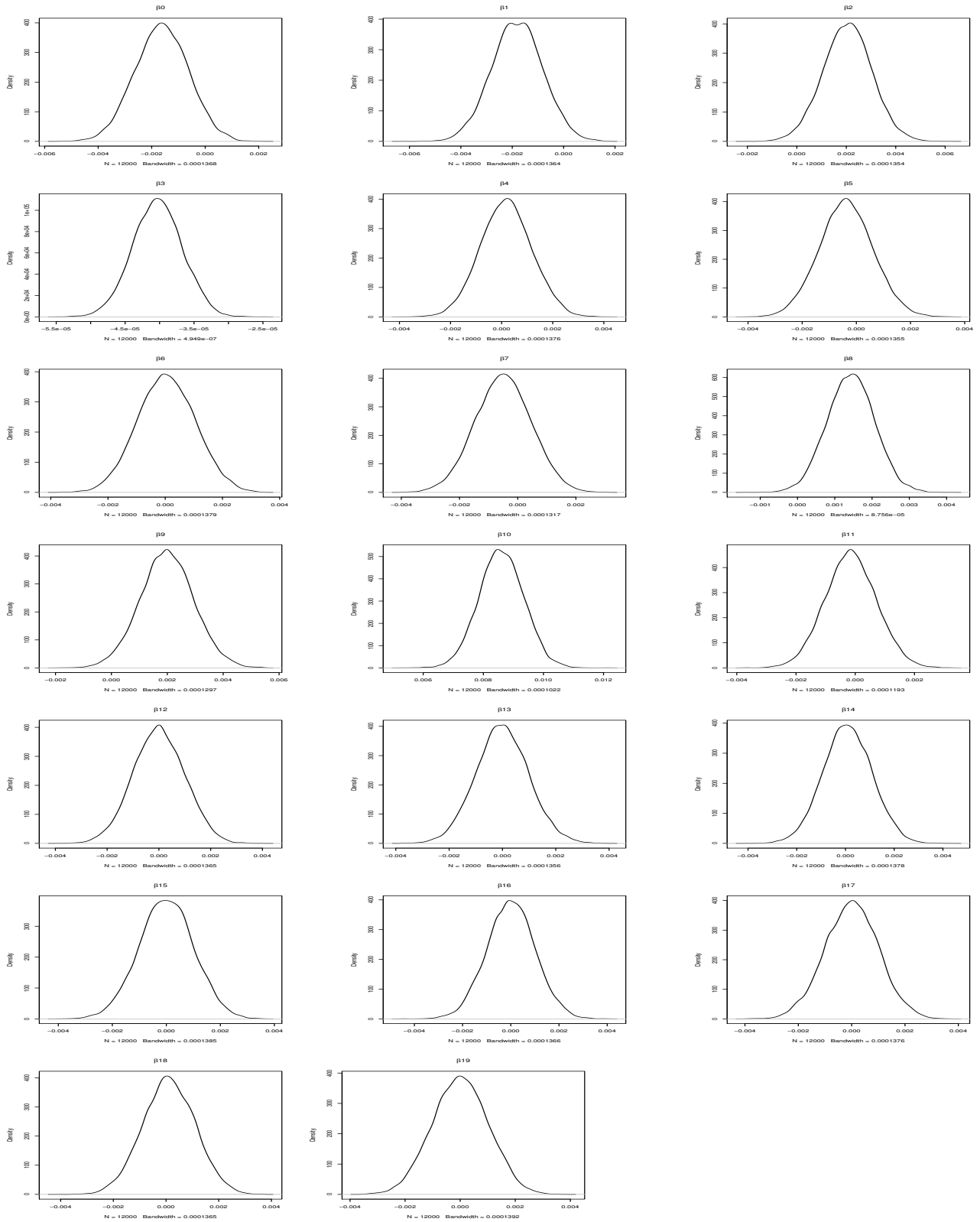


Figura 4: Densidade das cadeias para os parâmetros de β_0 a β_{19} .

Os gráficos das densidades são simétricos, se assemelhando a curva da distribuição normal, sugerindo que as distribuições *a priori*, ainda que não informativas, interferem no formato das distribuições marginais *a posteriori* obtidas.

Além das análises gráficas, também foram utilizados os critérios de Geweke e Raftery e Lewis para garantir que as séries estejam estacionárias.

Lembrando que o padrão do teste de Geweke compara os dez por cento inicial da cadeia de cada parâmetro, com cinquenta por cento final da própria cadeia, sendo tal padrão adotado neste trabalho. A partir deste critério, obtém-se uma estatística a qual pode ser comparada com os quantis da distribuição normal, ou valores críticos -1,96 e 1,96. Valores da estatística mais extremos do que estes indicaria falta de convergência.

Já o critério de Raftery e Lewis estima as iterações necessárias para convergência da cadeia a uma distribuição estacionária, sugestão de *burn in*, no qual a estatística do teste (fator de dependência) deve ser inferior a “cinco” para haver convergência.

Os resultados dos testes estão descritos na Tabela 3.

Tabela 3: Estatísticas dos testes de convergências das cadeias para cada parâmetro.

Parâmetros	Testes utilizados	
	Geweke	Raftery e Lewis
β_0	-0,82149	0,997
β_1	1,90128	1,010
β_2	0,66618	1,000
β_3	-0,06877	1,090
β_4	-0,18722	1,020
β_5	-0,18211	0,990
β_6	0,44771	1,010
β_7	-0,91100	1,000
β_8	-0,59167	1,070
β_9	0,89291	0,984
β_{10}	1,23456	1,030
β_{11}	0,82258	1,020
β_{12}	1,59509	0,997
β_{13}	-0,20366	1,010
β_{14}	0,08513	1,040
β_{15}	0,28450	0,964
β_{16}	-0,92463	0,984
β_{17}	-0,59490	1,000
β_{18}	0,52400	0,988
β_{19}	0,72911	0,990

Pelo critério de Raftery e Lewis, todas as cadeias convergiram, uma vez que os fatores de dependência foram todos menores que “cinco”.

A hipótese de convergência também não é rejeitada pelo critério de Geweke uma vez que a estatística deste teste está entre os valores de referência, citados anteriormente, para todas as cadeias.

Desta maneira conclui-se graficamente e por critérios considerados formais que as cadeias convergem, ou seja, são estacionárias.

Assim, pode-se determinar estimativas para os parâmetros, bem como utilizar os respectivos intervalos de credibilidade como critério inicial de seleção de modelos, porém trata-se apenas

de um indicativo. Deste modo, para decidir quais variáveis tem influência (permanecerão) no modelo aplica-se, como fator decisivo, o teste de hipóteses genuinamente bayesiano ou FBST.

Na Tabela 4, têm-se as estimativas dos parâmetros para o ajuste do modelo completo, juntamente com o intervalo de credibilidade, coeficiente de assimetria para as cadeias geradas pelas condicionais dos parâmetros e a estatística do valor-e para decidir qual variável deverá ser retirada do modelo, se houver necessidade.

Tabela 4: Estatísticas das cadeias dos parâmetros para o primeiro passo da seleção de variáveis

Parâmetros	Média	HPD Inferior	HPD Superior	Coeficiente de Assimetria	valor-e
β_0	0,000091	-0,001874	0,002050	-0,015346	0,913666
β_1	-0,001806	-0,003795	0,000077	0,015800	0,066166
β_2	0,002038	0,000079	0,003929	0,014276	0,041000
β_3	-0,000040	-0,000047	-0,000033	-0,005310	0,000000
β_4	0,000186	-0,001776	0,002072	0,040610	0,858166
β_5	-0,000363	-0,002403	0,001556	-0,014644	0,714500
β_6	0,000063	-0,001900	0,002037	-0,007945	0,952000
β_7	-0,000504	-0,002460	0,001308	-0,026207	0,606166
β_8	0,001437	0,000145	0,002608	-0,035072	0,022834
β_9	0,001995	0,000057	0,003832	-0,007278	0,046000
β_{10}	0,008589	0,007113	0,010019	0,015720	0,000000
β_{11}	-0,000142	-0,001850	0,001572	0,012665	0,881333
β_{12}	-0,000011	-0,001966	0,001983	-0,015751	0,997333
β_{13}	0,000017	-0,001920	0,002018	0,014310	0,997666
β_{14}	0,000095	-0,001959	0,001911	-0,011591	0,941166
β_{15}	0,000011	-0,001975	0,001956	-0,026639	0,991500
β_{16}	0,000008	-0,002054	0,001937	0,002862	0,986000
β_{17}	0,000100	-0,001801	0,002095	0,000313	0,929333
β_{18}	0,000010	-0,001841	0,001960	0,024031	0,994333
β_{19}	0,000012	-0,001943	0,001918	0,007398	0,971833

Com base nos primeiros resultados (Tabela 4) percebe-se que as cadeias são simétricas, já que o coeficiente de assimetria encontra-se dentro do intervalo especificado para não rejeição da hipótese de simetria.

Haja vista a simetria, o valor-e pode ser calculado multiplicando por dois a menor das proporções de valores da cadeia, a direita ou à esquerda do zero (hipótese nula). Utiliza-se o valor-e para decidir qual variável será excluída do modelo, neste sentido, nota-se que o maior valor-e obtido foi para a cadeia do parâmetro β_{13} (valor-e = 0,997666) referente a variável binária “ x_{13} ”, que especifica quais pessoas são casadas. Quanto maior o valor-e, mais evidências se tem a favor da hipótese de nulidade do parâmetro.

Retirando a variável “ x_{13} ”, aplica-se novamente o ajuste do modelo para as demais variáveis, tratando como o segundo passo para seleção de variáveis a partir do modelo completo.

Novamente verificou-se que as cadeias *a posteriori* possuem séries estacionárias pelos critérios gráficos e numéricos. Assim pode-se avaliar outra vez as estatísticas de assimetria e valor-e, para avaliar a necessidade de exclusão ou não de outra variável do modelo.

A estatísticas obtidas seguem de acordo com a Tabela 5.

Tabela 5: Estatísticas das cadeias dos parâmetros para o segundo passo da seleção de variáveis

Parâmetros	Média	HPD Inferior	HPD Superior	Coefficiente de Assimetria	valor-e
β_0	0,000110	-0,001772	0,002133	0,016894	0,926166
β_1	-0,001807	-0,003650	0,000185	-0,008425	0,069166
β_2	0,002042	0,000221	0,003989	0,023403	0,035833
β_3	-0,000040	-0,000047	-0,000033	-0,030093	0,000000
β_4	0,000173	-0,001789	0,002105	-0,040518	0,860000
β_5	-0,000391	-0,002362	0,001570	-0,028989	0,707000
β_6	0,000048	-0,001885	0,002087	0,022859	0,925833
β_7	-0,000489	-0,002405	0,001365	0,010152	0,594333
β_8	0,001435	0,000193	0,002634	0,005999	0,023833
β_9	0,001973	0,000056	0,003905	0,047745	0,038666
β_{10}	0,008602	0,007154	0,010046	0,021815	0,000000
β_{11}	-0,000153	-0,001910	0,001514	0,019162	0,869000
β_{12}	-0,000011	-0,001982	0,001876	-0,049011	0,996500
β_{14}	0,000017	-0,001967	0,001917	-0,032057	0,921500
β_{15}	0,000006	-0,001955	0,001973	-0,041461	0,997833
β_{16}	0,000009	-0,001922	0,001987	0,023178	0,990166
β_{17}	0,000102	-0,001838	0,002057	-0,000794	0,910000
β_{18}	0,000004	-0,001993	0,001905	0,022503	0,982666
β_{19}	0,000021	-0,001944	0,001993	0,009835	0,980500

No segundo passo (Tabela 5) para seleção de variáveis, percebe-se que as cadeias das distribuições *a posteriori* são simétricas, por meio dos coeficientes de assimetria, os quais estão dentro do intervalo de $-0,2$ a $0,2$.

Pelo valor-e, nota-se que a maior estimativa (valor-e = 0,997833), que não é considerada influente a 5% de probabilidade, foi para o parâmetro β_{15} referente a variável x_{15} . Sendo assim a variável relativa a indivíduos que possuem ensino fundamental incompleto será excluída do modelo.

Seguindo com o mesmo raciocínio, na Tabela 6 é possível visualizar os resultados para o terceiro passo realizado.

Tabela 6: Estatísticas das cadeias dos parâmetros para o terceiro passo da seleção de variáveis

Parâmetros	Média	HPD Inferior	HPD Superior	Coefficiente de Assimetria	valor-e
β_0	0,000100	-0,001856	0,001989	-0,017677	0,923667
β_1	-0,001794	-0,003786	0,000109	0,017246	0,072167
β_2	0,002029	0,000013	0,003872	-0,003034	0,040000
β_3	-0,000040	-0,000047	-0,000033	-0,020354	0,000000
β_4	0,000194	-0,001762	0,002118	-0,013388	0,842833
β_5	-0,000380	-0,002326	0,001532	0,043025	0,701000
β_6	0,000066	-0,001879	0,001996	0,040685	0,951167
β_7	-0,000505	-0,002312	0,001414	-0,025993	0,609000
β_8	0,001427	0,000163	0,002625	-0,008343	0,024000
β_9	0,001971	0,000087	0,003855	-0,003657	0,041500
β_{10}	0,008595	0,007151	0,010069	-0,016086	0,000000
β_{11}	-0,000136	-0,001928	0,001545	-0,004854	0,879833
β_{12}	0,000007	-0,001922	0,001972	-0,017200	0,999667
β_{14}	0,000106	-0,001827	0,002092	0,020520	0,911833
β_{16}	0,000008	-0,002003	0,001915	-0,001308	0,988667
β_{17}	0,000097	-0,001920	0,002028	-0,006949	0,922833
β_{18}	0,000028	-0,001916	0,001998	0,003730	0,978833
β_{19}	0,000018	-0,001957	0,001973	-0,009721	0,983167

Para o terceiro passo (Tabela 6), a variável x_{12} referente a taxa média de conversão, possui maior valor-e (valor-e = 0,999667), sendo assim, excluída do modelo.

No quarto passo, de acordo com os dados presentes na Tabela 7, os resultados do ajuste foram:

Tabela 7: Estatísticas das cadeias dos parâmetros para o quarto passo da seleção de variáveis

Parâmetros	Média	HPD Inferior	HPD Superior	Coefficiente de Assimetria	valor-e
β_0	0,000110	-0,001870	0,002061	0,000281	0,912000
β_1	-0,001802	-0,003707	0,000128	0,006824	0,067333
β_2	0,002036	0,000070	0,003932	-0,02884	0,040000
β_3	-0,000040	-0,000047	-0,000033	0,016609	0,000000
β_4	0,000185	-0,001793	0,002132	-0,021841	0,857000
β_5	-0,000369	-0,002236	0,001651	-0,013586	0,713333
β_6	0,000068	-0,001830	0,002044	-0,007020	0,950833
β_7	-0,000505	-0,002322	0,001397	-0,000605	0,592167
β_8	0,001438	0,000245	0,002679	-0,010187	0,021333
β_9	0,001967	0,000167	0,003905	0,007899	0,038833
β_{10}	0,008605	0,007115	0,010025	-0,008274	0,000000
β_{11}	-0,000137	-0,001866	0,001550	-0,003113	0,877167
β_{14}	0,000090	-0,001950	0,001988	-0,013336	0,934500
β_{16}	0,000001	-0,001988	0,001964	-0,018375	0,999833
β_{17}	0,000107	-0,001928	0,001984	-0,044076	0,906000
β_{18}	0,000023	-0,001983	0,001963	-0,006550	0,986167
β_{19}	0,000015	-0,001933	0,001984	0,034213	0,984000

No quarto passo (Tabela 7), a variável x_{16} referente a pessoas que possuem ensino fundamental completo, apresentou maior valor-e (valor-e = 0,999833), sendo por tanto, excluída do modelo.

Passando ao quinto passo, na Tabela 8 tem-se os resultados:

Tabela 8: Estatísticas das cadeias dos parâmetros para o quinto passo da seleção de variáveis

Parâmetros	Média	HPD Inferior	HPD Superior	Coefficiente de Assimetria	valor-e
β_0	0,000096	-0,001867	0,002080	-0,013889	0,925667
β_1	-0,001804	-0,003732	0,000110	-0,011554	0,066500
β_2	0,002026	0,000152	0,003981	-0,012483	0,040667
β_3	-0,000040	-0,000047	-0,000033	-0,041748	0,000000
β_4	0,000173	-0,001786	0,002090	-0,009172	0,856667
β_5	-0,000382	-0,002280	0,001618	-0,020314	0,691167
β_6	0,000054	-0,001959	0,001976	0,013377	0,958000
β_7	-0,000488	-0,002303	0,001384	0,018348	0,599000
β_8	0,001420	0,000247	0,002705	-0,028729	0,024667
β_9	0,001979	0,000095	0,003920	0,012331	0,043167
β_{10}	0,008594	0,007191	0,010108	-0,052494	0,000000
β_{11}	-0,000153	-0,001858	0,001551	0,032933	0,853167
β_{14}	0,000104	-0,001861	0,002057	0,015054	0,915333
β_{17}	0,000096	-0,001824	0,002062	-0,012109	0,921833
β_{18}	0,000023	-0,001931	0,002011	0,039346	0,984333
β_{19}	0,000020	-0,002018	0,001876	-0,026430	0,980833

Neste passo (Tabela 8), a variável x_{18} , referente ao gênero dos indivíduos é retirada do modelo por apresentar o maior valor-e (valor-e = 0,984333).

No sexto passo, pela Tabela 9, temos os resultados:

Tabela 9: Estatísticas das cadeias dos parâmetros para o sexto passo da seleção de variáveis

Parâmetros	Média	HPD Inferior	HPD Superior	Coefficiente de Assimetria	valor-e
β_0	0,000091	-0,001869	0,002044	0,018599	0,936000
β_1	-0,001804	-0,003735	0,000131	-0,001950	0,068167
β_2	0,002046	0,000147	0,003984	0,028835	0,036167
β_3	-0,000040	-0,000047	-0,000033	-0,032112	0,000000
β_4	0,000172	-0,001855	0,002066	0,001456	0,861667
β_5	-0,000360	-0,002377	0,001575	0,017198	0,720500
β_6	0,000068	-0,001778	0,002119	-0,016789	0,949333
β_7	-0,000500	-0,002343	0,001368	-0,022108	0,597167
β_8	0,001424	0,000167	0,002628	-0,028187	0,023667
β_9	0,001986	0,000131	0,003887	0,007173	0,038000
β_{10}	0,008596	0,007171	0,010048	0,018105	0,000000
β_{11}	-0,000147	-0,001897	0,001510	-0,040785	0,873167
β_{14}	0,000099	-0,001870	0,001995	0,003839	0,925500
β_{17}	0,000094	-0,001844	0,002022	0,028626	0,939333
β_{19}	0,000039	-0,001893	0,001985	0,016939	0,968833

Neste caso (Tabela 9), a variável com maior valor-e (valor-e = 0,968833) excluída do modelo é x_{19} , que se refere às pessoas que possuem algum dependente econômico.

Para o sétimo passo, tem-se na Tabela 10 os seguintes resultados:

Tabela 10: Estatísticas das cadeias dos parâmetros para o sétimo passo da seleção de variáveis

Parâmetros	Média	HPD Inferior	HPD Superior	Coefficiente de Assimetria	valor-e
β_0	0,000109	-0,001862	0,002052	-0,032719	0,910000
β_1	-0,001801	-0,003678	0,000130	-0,015500	0,063000
β_2	0,002023	0,000121	0,003959	-0,005789	0,043333
β_3	-0,000040	-0,000047	-0,000033	-0,028917	0,000000
β_4	0,000176	-0,001828	0,002061	-0,000364	0,849500
β_5	-0,000370	-0,002265	0,001672	-0,003217	0,714667
β_6	0,000071	-0,001888	0,002022	-0,002258	0,951500
β_7	-0,000499	-0,002382	0,001379	0,019254	0,605000
β_8	0,001438	0,000218	0,002651	-0,006889	0,020833
β_9	0,001971	0,000083	0,003909	-0,028601	0,045833
β_{10}	0,008604	0,007134	0,010074	0,020875	0,000000
β_{11}	-0,000150	-0,001823	0,001614	-0,000650	0,872833
β_{14}	0,000096	-0,001861	0,002093	0,000125	0,922333
β_{17}	0,000098	-0,001938	0,002027	-0,000951	0,918500

Assim, a variável de maior valor-e (valor-e = 0,951500) a ser excluída no sétimo passo (Tabela 10) foi x_6 , referente ao índice do vendedor.

Partindo para o oitavo passo, obteve-se os resultados de acordo com a Tabela 11.

Tabela 11: Estatísticas das cadeias dos parâmetros para o oitavo passo da seleção de variáveis

Parâmetros	Média	HPD Inferior	HPD Superior	Coefficiente de Assimetria	valor-e
β_0	0,000103	-0,001770	0,002086	-0,024405	0,918000
β_1	-0,001816	-0,003731	0,000130	-0,012192	0,064000
β_2	0,002019	0,000174	0,003909	-0,014522	0,038833
β_3	-0,000040	-0,000047	-0,000033	-0,022265	0,000000
β_4	0,000174	-0,001753	0,002129	0,006241	0,851000
β_5	-0,000385	-0,002317	0,001544	-0,011968	0,693667
β_7	0,000061	-0,001874	0,002027	-0,001997	0,948833
β_8	0,001378	0,000150	0,002598	0,005850	0,028167
β_9	0,001981	0,000149	0,003944	-0,007184	0,043000
β_{10}	0,008610	0,007142	0,010023	0,040081	0,000000
β_{11}	-0,000135	-0,001780	0,001673	0,023761	0,864667
β_{14}	0,000096	-0,001867	0,002010	-0,010660	0,919000
β_{17}	0,000104	-0,001895	0,002024	-0,031802	0,911333

No oitavo passo (Tabela 11), a variável x_7 referente ao prêmio médio pago por economia de flexível, apresentou o maior valor-e (valor-e = 0,948833), sendo então retirada do modelo.

Na Tabela 12, tem-se os resultados para o nono passo da seleção de variáveis.

Tabela 12: Estatísticas das cadeias dos parâmetros para o nono passo da seleção de variáveis

Parâmetros	Média	HPD Inferior	HPD Superior	Coefficiente de Assimetria	valor-e
β_0	0,000091	-0,001823	0,002043	-0,044726	0,925833
β_1	-0,001802	-0,003707	0,000135	0,002759	0,065333
β_2	0,002028	0,000080	0,003921	-0,033168	0,039500
β_3	-0,000040	-0,000047	-0,000033	-0,029282	0,000000
β_4	0,000157	-0,001823	0,002055	-0,020138	0,874667
β_5	-0,000364	-0,002342	0,001517	0,014435	0,718667
β_8	0,001381	0,000162	0,002595	-0,001211	0,026000
β_9	0,001983	0,000103	0,003911	0,023681	0,041000
β_{10}	0,008609	0,007257	0,010114	0,023639	0,000000
β_{11}	-0,000133	-0,001879	0,001619	0,012613	0,877833
β_{14}	0,000098	-0,001812	0,002074	-0,005916	0,925500
β_{17}	0,000103	-0,001783	0,002117	-0,018138	0,911000

Feito o nono passo (Tabela 12), o maior valor-e (valor-e = 0,925833) foi obtido para o intercepto x_0 , de modo que este foi removido do modelo.

Os resultados obtidos para o décimo passo, são encontrados na Tabela 13.

Tabela 13: Estatísticas das cadeias dos parâmetros para o décimo passo da seleção de variáveis

Parâmetros	Média	HPD Inferior	HPD Superior	Coefficiente de Assimetria	valor-e
β_1	-0,001798	-0,003724	0,000142	-0,003114	,065000
β_2	0,002038	0,000126	0,003923	-0,040582	,038333
β_3	-0,000040	-0,000047	-0,000033	-0,047359	,000000
β_4	0,000167	-0,001820	0,002047	-0,010564	0,858333
β_5	-0,000361	-0,002396	0,001471	-0,008782	0,718167
β_8	0,001385	0,000166	0,002610	0,008338	0,025333
β_9	0,001980	0,000030	0,003796	0,005346	0,037167
β_{10}	0,008608	0,007143	0,010048	0,012607	0,000000
β_{11}	-0,000140	-0,001895	0,001519	-0,007836	0,877500
β_{14}	0,000084	-0,001815	0,002020	-0,038089	0,923000
β_{17}	0,000111	-0,001789	0,002137	-0,004283	0,911500

A variável de maior valor-e (valor-e = 0,923000) a ser excluída no décimo passo é x_{14} (Tabela 13), que se refere aos indivíduos que são solteiros.

Realizada a rotina para o décimo primeiro passo, conforme a Tabela 14, os resultados foram:

Tabela 14: Estatísticas das cadeias dos parâmetros para o décimo primeiro passo da seleção de variáveis

Parâmetros	Média	HPD Inferior	HPD Superior	Coefficiente de Assimetria	valor-e
β_1	-0,001835	-0,003820	0,000040	0,007116	0,062500
β_2	0,002068	0,000194	0,004004	0,023086	0,032500
β_3	-0,000040	-0,000047	-0,000034	-0,053222	0,000000
β_4	0,000173	-0,001765	0,002107	-0,021130	0,855167
β_5	-0,000379	-0,002316	0,001543	0,009594	0,700500
β_8	0,001383	0,000143	0,002583	0,025967	0,024500
β_9	0,001971	0,000063	0,003915	-0,004371	0,046500
β_{10}	0,008612	0,007206	0,010072	-0,028892	0,000000
β_{11}	-0,000139	-0,001907	0,001567	0,027384	0,873833
β_{17}	0,000083	-0,001817	0,002084	0,050357	0,947667

Percebe-se que o maior valor-e (valor-e = 0,947667) obtido na Tabela 14 foi para o parâmetro referente a variável x_{17} , que se trata de indivíduos que possuem ensino médio completo. Sendo assim, esta foi excluída do modelo.

No décimo segundo passo, os resultados seguem na Tabela 15.

Tabela 15: Estatísticas das cadeias dos parâmetros para o décimo segundo passo da seleção de variáveis

Parâmetros	Média	HPD Inferior	HPD Superior	Coefficiente de Assimetria	valor-e
β_1	-0,0018141	-0,0038032	0,0000715	-0,0102103	0,0643333
β_2	0,0020481	0,0000728	0,0038731	-0,0160692	0,0358333
β_3	-0,0000404	-0,0000474	-0,0000336	-0,0181052	0,0000000
β_4	0,0001639	-0,0018309	0,0020223	-0,0242482	0,8688333
β_5	-0,0003785	-0,0022807	0,0015901	0,0215062	0,6938333
β_8	0,0013809	0,0001733	0,0026178	0,0094735	0,0281667
β_9	0,0019796	0,0000358	0,0038245	0,0109934	0,0431667
β_{10}	0,0086162	0,0072025	0,0101074	0,0558875	0,0000000
β_{11}	-0,0001415	-0,0018463	0,0015920	0,0125561	0,8708333

Assim, a variável que apresentou maior valor-e (valor-e = 0,870833) no décimo segundo passo (Tabela 15) foi x_{11} , referente a nota média de qualidade, sendo assim, excluída do modelo. Realizado o décimo terceiro passo da seleção de variáveis, de acordo com a Tabela 16, os resultados foram:

Tabela 16: Estatísticas das cadeias dos parâmetros para o décimo terceiro passo da seleção de variáveis

Parâmetros	Média	HPD Inferior	HPD Superior	Coefficiente de Assimetria	valor-e
β_1	-0,00181	-0,0038271	0,0000789	0,0040819	0,066833
β_2	0,002035	0,0000456	0,0039170	0,0269513	0,039000
β_3	-0,000045	-0,0000476	-0,0000336	-0,0102003	0,000000
β_4	0,000152	-0,0017560	0,0020972	0,0288361	0,879500
β_5	-0,00037	-0,0023030	0,0016448	0,0353680	0,717333
β_8	0,001387	0,0001138	0,0025768	-0,0437746	0,028667
β_9	0,001972	0,000119	0,003866	0,006759	0,038000
β_{10}	0,008599	0,007188	0,010055	-0,005876	0,000000

A variável x_4 referente a quantidade média de pedidos, apresentou o maior valor-e (valor-e = 0,879500) não influente a 5% de probabilidade (Tabela 16), sendo então, excluída do modelo.

No décimo quarto passo representado na Tabela 17, os resultados obtidos foram:

Tabela 17: Estatísticas das cadeias dos parâmetros para o décimo quarto passo da seleção de variáveis

Parâmetros	Média	HPD Inferior	HPD Superior	Coefficiente de Assimetria	valor-e
β_1	-0,001813	-0,003695	0,000189	0,003620	0,067667
β_2	0,002039	0,000129	0,003983	0,006807	0,037667
β_3	-0,000040	-0,000047	-0,000034	-0,009531	0,000000
β_5	-0,000389	-0,002436	0,001449	-0,002023	0,698667
β_8	0,001391	0,000152	0,002586	-0,033463	0,026500
β_9	0,001972	0,000045	0,003850	-0,000185	0,040833
β_{10}	0,008607	0,007214	0,010080	-0,012723	0,000000

No décimo quarto passo (Tabela 17), o maior valor-e obtido (valor-e = 0,698667) foi para a variável x_5 , que retrata a quantidade de SKU's realizado pelos vendedores, de forma que esta foi eliminada do modelo.

Seguindo, na Tabela 18 tem-se os resultados referentes ao décimo quinto passo da seleção de variáveis.

Tabela 18: Estatísticas das cadeias dos parâmetros para o décimo quinto passo da seleção de variáveis

Parâmetros	Média	HPD Inferior	HPD Superior	Coefficiente de Assimetria	valor-e
β_1	-0,001801	-0,003725	0,000127	0,067500	0,067500
β_2	0,002035	0,000195	0,004008	0,036833	0,036833
β_3	-0,000040	-0,000047	-0,000034	0,000000	0,000000
β_8	0,001388	0,000199	0,002604	0,021667	0,021667
β_9	0,001973	0,000103	0,003870	0,041333	0,041333
β_{10}	0,008594	0,007190	0,010042	0,000000	0,000000

Para o décimo quinto passo da seleção de variáveis (Tabela 18), x_1 que trata do tempo de casa dos vendedores, apresentou o maior valor-e não significativo dentre as demais variáveis (valor-e = 0,067500), sendo retirada do modelo.

Prosseguindo, o décimo sexto passo resultou na Tabela 19 da seleção de variáveis.

Tabela 19: Estatísticas das cadeias dos parâmetros para o décimo sexto passo da seleção de variáveis

Parâmetros	Média	HPD Inferior	HPD Superior	Coefficiente de Assimetria	valor-e
β_2	0,002035	0,000071	0,003938	-0,007621	0,038166
β_3	-0,000041	-0,000048	-0,000035	-0,041701	0,000000
β_8	0,001391	0,000230	0,002630	0,018900	0,021000
β_9	0,001976	0,000074	0,003882	0,000345	0,043334
β_{10}	0,008586	0,007162	0,010070	-0,013263	0,000000

Concluiria-se este como sendo o último passo da seleção de variáveis (Tabela 19), no qual as cadeias *a posteriori* para os parâmetros β_2 , β_3 , β_8 , β_9 e β_{10} foram simétricas e apresentaram valor-e com probabilidade de menor do que 5%.

Porém, do ponto de vista de negócio, não faz sentido que a variável x_8 , referente ao prêmio pago aos vendedores, apresente parâmetro positivo, pois sua interpretação seria que ao aumentar o valor recebido de prêmio, também aumentaria as chances de desligamento em menos de um ano trabalhado.

Ainda, identificou-se que x_8 e x_3 apresentaram correlação de 0,88, considerada forte. Assim, as duas variáveis juntas, no modelo, geram fortes indícios de que existe o efeito de multicolinearidade entre elas, tal efeito acontece quando duas ou mais variáveis acrescentam informações similares ao modelo ajustado, de forma que a entrada de β_8 estaria atuando para ponderar a estimativa de β_3 .

Observando graficamente, conforme a Figura (5), nota-se um comportamento linear positivo entre as variáveis x_3 e x_8 , indicando novamente a existência de multicolinearidade no modelo.

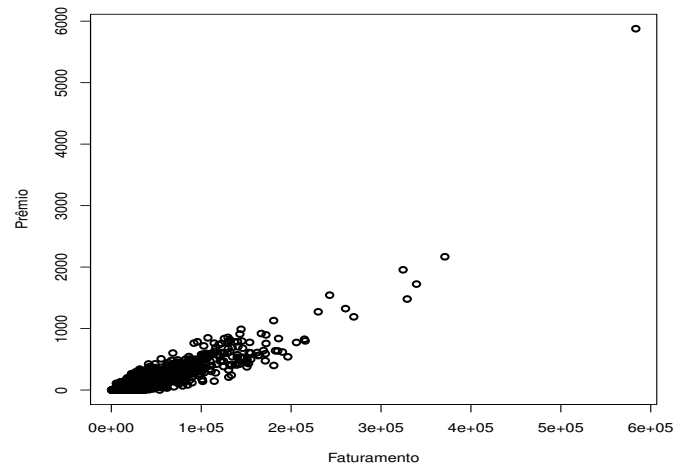


Figura 5: Plot entre as variáveis x_3 a x_8 para confirmação da existência de multicolineariedade.

Com isso, realizou-se novamente o ajuste do modelo, agora sem a variável x_8 , uma vez que este obteve melhor assertividade do que se fosse retirada x_3 , obtendo os resultados conforme a Tabela 20.

Tabela 20: Estatísticas das cadeias dos parâmetros para o novo modelo sem x_8 (décimo sétimo passo)

Parâmetros	Média	HPD Inferior	HPD Superior	Coefficiente de Assimetria	valor-e
β_2	0,001959	0,000075	0,003882	-0,007621	0,043334
β_3	-0,000035	-0,000039	-0,000030	-0,041701	0,000000
β_9	0,001894	0,000056	0,003719	0,003345	0,051833
β_{10}	0,008401	0,007194	0,010070	-0,013263	0,000000

Com a retirada de x_8 do modelo a variável x_9 , referente a média de chamadas efetivas realizadas por cada vendedor em média, tornou-se não influente para o modelo 5% de probabilidade (valor-e = 0,051833), desta forma foi necessário o reajuste do modelo sem a mesma.

Com reajuste, obteve-se os resultados do décimo oitavo passo realizado, de acordo com a Tabela 21.

Tabela 21: Estatísticas das cadeias dos parâmetros para o modelo final com x_2 , x_3 e x_{10} (décimo oitavo passo)

Parâmetros	Média	HPD Inferior	HPD Superior	Coefficiente de Assimetria	valor-e
β_2	0,002032	0,000157	0,003947	0,016631	0,034666
β_3	-0,000035	-0,000039	-0,000030	-0,015792	0,000000
β_{10}	0,008637	0,007190	0,009989	0,006331	0,000000

Assim, sem mais modificações, as variáveis que permaneceram no modelo foram x_2 referente a idade, x_3 atribuído a faturamento médio e x_{10} que retrata o tempo médio em chamadas.

Em termos matemáticos, o modelo obtido para o cálculo das probabilidades dos colaboradores do setor de telemarketing desligarem da empresa com menos de um ano de trabalhado (caracterizando prejuízo), é dado por:

$$P(y = 1|\beta_1, \beta_2, \dots, \beta_{19}) = \frac{e^{0,002032x_2 - 0,000035x_3 + 0,008637x_{10}}}{1 + e^{0,002032x_2 - 0,000035x_3 + 0,008637x_{10}}} \quad (14)$$

Com o modelo estimado em mãos podemos obter o "*odds ratio*"-OR (razão de chances), para determinar o quão influente cada parâmetro é para o modelo proposto, ou seja o quanto a informação dos mesmos trazem ao incrementar-se valores nas variáveis explicativas.

Para o parâmetro β_2 , referente a variável idade, tem-se OR igual a 1,002034, ou seja, para cada um ano de idade que o vendedor completa a mais, aumenta-se em 0,2034% a probabilidade do mesmo permanecer menos que um ano em serviço. Se forem tomados dois indivíduos, um com 20 anos de idade e outro com 30, sendo que ambos apresentassem mesmo faturamento e mesmo tempo médio em chamada, o mais velho teria 2,053% de chance a mais de sair da empresa em menos de um ano.

Tratando o parâmetro β_3 , referente ao faturamento médio, tem-se OR de 0,99959, de forma que para cada um real a mais na média de faturamento, reduz-se em 0,003499% as chances do vendedor sair da empresa em menos de um ano. Considerando a existência de dois vendedores, um com média de 50.000,00 reais e outro com média de 100.000,00 reais (diferença bastante comum), sendo que ambos possuísem mesma idade e mesmo tempo médio em atendimento, as chances daquele com menor desempenho sair da empresa em menos de um ano trabalhado seria 5,75 vezes maior.

Por fim, para o parâmetro β_{10} , que trata a influência do tempo médio em chamada no modelo, tem-se OR igual a 1,008674, de modo que a cada segundo a mais, em média, falado pelos

vendedores em suas ligações, aumenta em 0,8674% as chances de serem desligados da empresa antes de completarem um ano de companhia. Se hipoteticamente existisse dois operadores, um com média de 60 segundo em atendimento e outro com média de 80 segundos falados, sendo que ambos possuíssem mesmo faturamento médio e mesma idade, as chances daquele que possui maior tempo não permanecer na empresa mais que um ano seria 18,8% maior.

Para o modelo em geral, em termos percentuais, do total de classificações realizadas pelo modelo (1540), 1098 foram realizadas de forma correta, logo obteve-se 73,5% de assertividade, valor-este considerado bom pelos administradores da setor de telemarketing. No conceito de “assertividade” utilizado, foram consideradas as classificações corretas do modelo, ou seja, quando o evento contido na variável resposta do indivíduo é “zero” e no modelo gerado obteve-se probabilidade abaixo de 0,5 ou 50%, ou quando o evento do indivíduo é “um” e no modelo sua classificação é maior que 0,5 ou 50%.

Esta avaliação pode ser vista pelo gráfico do cruzamento entre a variável resposta, contendo os eventos ocorridos até a data de coleta das informações, e as classificações realizadas pelo modelo conforme a Figura 6

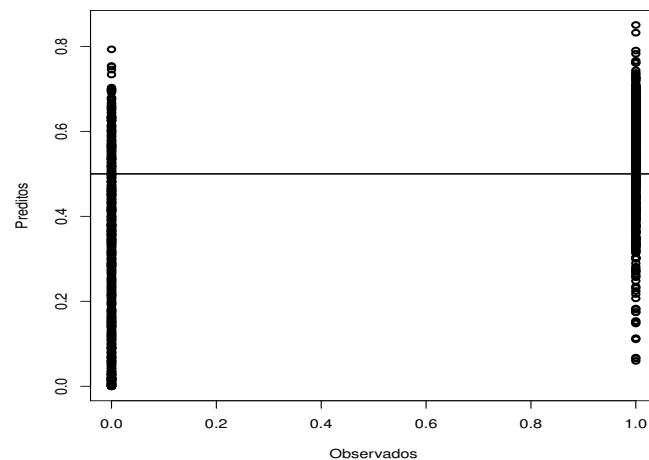


Figura 6: Plot dos valores preditos contra os observados.

Na Figura 6 é nítida a diferença entre as probabilidades entre os grupos de desligados ou que estão ativos com menos de um ano trabalhado (1) e os que permaneceram ou estão na empresa a mais de um ano (0).

Quando realizado ajuste do modelo com a variável prêmio médio (x_8) em vez de faturamento médio (x_3), a assertividade do modelo foi de 61,4%, com 947 classificações corretas, por isso a ajuste com a variável faturamento médio foi mantida após a identificação da existência de multicolinearidade no modelo

Após o último passo da seleção de variáveis na Tabela 21, verifica-se que as variáveis que restaram também são as que apresentaram diferenças de valores entre os grupos da variável resposta do modelo presentes nos gráficos da Figura 2. No entanto o estudo da relação entre as preditoras e a variável resposta, por meios gráficos, é subjetiva e deve ser utilizada apenas para uma exploração prévia dos dados.

5 conclusões

Por meio deste trabalho foi possível obter um bom modelo logístico bayesiano, que além de gerar uma alta proporção de acertos nas previsões, permitiu identificar que fatores como idade, tempo médio em chamada e faturamento são relevantes para mapear vendedores propensos ao desligamento, sendo que este último fator mostrou ser o mais impactante no cálculo das probabilidades.

O modelo obtido não é único. Outros modelos são possíveis se a decisão for pela permanência de variáveis fundamentais ao negócio.

O valor-e, é uma medida probabilística da evidência favorável à hipótese nula possível mesmo quando a distribuição marginal *a posteriori* for empírica.

A probabilidade de referência para decidir sobre a exclusão de uma variável deve ser determinada conforme o rigor adotado na seleção das variáveis do modelo. Mais rigor implicaria em menos variáveis selecionadas e menor valor de referência para o valor-e. O valor de 5% adotado neste trabalho foi arbitrário.

Haja vista o enfoque bayesiano, as estimativas dos parâmetros do modelo final obtido neste trabalho podem ser utilizadas como informação prévia em análises futuras.

6 Referências

Referências

- [1] ABT ASSOCIAÇÃO BRASILEIRA DE TELESERVIÇOS (ABT).: *Institucional*. Disponível em: <<http://www.abt.org.br/institucional.asp?nanner=ABT>> Acesso em jul. 2017.
- [2] BOX, G.E.P., TIÃO, G.C.: *Bayesian inference in statistical analysis*. New York: Addison Wesley, 588p., 1973.
- [3] BRASIL, G.H.: *Estatística Bayesiana: alguns aspectos Teóricos e Práticos*, in: reunião Regional de Estatística, 22., 1991, Recife. Minicurso. Recife, UFPE, Apostila, 1991.
- [4] BUSSAB, W., MORETTIN, P.: *A estatística básica*. 5. ed. São Paulo:Saraiva, 2002.
- [5] CHIAVENATO, I.: *Recursos Humanos:edição compacta*. 5 edição. São Paulo: Atlas, 1998.
- [6] COSTA, L.B.: *Bayesiana do modelo dinâmico para um vetor de séries temporais usando distribuições elípticas*, São Paulo. 2008. 91f. Tese(Doutorado em Estatística)-Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.
- [7] DREISEITL, S., OHNO-MACHADO, L.: *Logistic regression and neural network*. Classification models: a methodology review. Journal of Biomedical Informatics, V.35, n.5-6, 352-359p., 2002.
- [8] EBRAHIM, G.J.: *Análise multivariada*. Tradução: Antônio Carlos Figueira et al. 160p. Recife:IMIP, 2001.
- [9] GELFAND, A.E., HILLS, S.E., SMITH, A.M.F.: *Illustration of bayesian inference in normal data models using Gibbs sampling*. J. Am. Statist., 85, 972-985, 1990.
- [10] GEWEKE, J.: *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. In J.M. Bernardo, J.O. Berger, A.P., & A.F.M. Smith (Eds.), *Bayesian Statistics 4* (pp. 169-193). New York: Oxford University Press, 1992.
- [11] GUIMARÃES, I.A.: *Construção e avaliação de uma regra de reconhecimento e classificação de clientes de uma instituição financeira com base na análise multivariada*. 2000. 87f. (dissertação Mestrado)- Métodos Numéricos em Engenharia, Universidade federal do Paraná, Curitiba, 2000.

- [12] HOSMER, D.W., LEMESHOW, S.: *Applied Logistic Regression*. 307p. New York: John Wiley, 1989.
- [13] JUBILATO J.: *Rotatividade de pessoal: custos e índice ideias*. Disponível em <<http://www.rh.com.br/Portal/RelacaoTrabalhista/Artigo/4980/rotatividadedepessoal-custos-e-indice-ideias.html>> Acesso em 08 dez 2017
- [14] LAURETTO, M.S.: *Seleção de modelos através de um teste genuinamente bayesiano: Misturas de normais multivariadas e hipóteses separadas*. Cap. 2, (pp.5-7). Universidade de São Paulo, São Paulo, 2007.
- [15] LOPES, L.F.D.: *Apostila Estatística*. Cap. 4, (pp.47). Universidade de Mato Grosso do Sul, Campo Grande, 2003
- [16] MCINTURFF, P., JOHNSON, W.O., COWLING, D., GARDNER I.A.: *Modelling risk when binary outcomes are subect to error*. University of California, Los Angeles, 2004.
- [17] PEREIRA, C.A.B., Stern, J.M.: *Evidence and credibility: full bayesian significance test of precise hypothesis*. Entropy, 1:99-110, 1999.
- [18] PLUMMER, M.: *JAGS version 3.1 user manual*. International Agency for Research on cancer. 12,16, 2011.
- [19] PLUMMER, M.: *Rjags: Bayesian graphical models using MCMC*. R package 3-10. URL: <http://CRAN.R-project.org/package=rjags>, 2013.
- [20] POMI, R.M.: *A importância da gestão do turnover*. Disponível em: <<http://www.rh.com.br/ler.php?cod=3998>> Acesso em 10 jul 2017.
- [21] R Core Team.: *R: A Language and Enviroment for Statistical Computing*. R Foudation for Statistical Computing. Vienna, Austria. 12,15, 2017.
- [22] RAFTERY, A.L., LEWIS, S.: *How many iteratios in the Gibbs sampler?* In J.M. Bernardo, J.O. Berger, A.P., & A.F.M. Smith (Eds.), *Bayesian Statistics 4* (pp. 763-773). New York: Oxford University Press, 1992.
- [23] ROBBINS, S.P.: *Comportamento Organizacional*. 11. ed.-São Pulo: Pearson Prentice Hall, 2005.
- [24] ROCHA, E.C., AGUIRELLA, F.,: *Rotatividade em call center: para além de indicadores, sem chamada a ação*. 373p. Volume 25. São Paulo, 2016.

- [25] ROSÁRIO, C.C. *Rotatividade de pessoal*. (2006) Disponível em <http://www.atituderh.com/artigo.aspx?Codigo=2177> Acesso em 10 jul 2017.
- [26] SANTANA F.A., POLO, F.J.V., PÉRES, J.C.: *Herramientas paar la investigación bio-médica: La perspectiva bayesiana*. Med Clin. Barcelona, V.119, n.7, 265-269p., 2002.
- [27] SILVA, A.L.C., SUÁREZ, G.R.: *Qué es la inferência bayesiana?* Jano EMC. Viernes, V.58, n.1338, 65-66p., 2000.
- [28] VALLEJOS, J.A., MCKINNON, S.D.: *Logistic regression and neural network. Classification of seismic records*. International journal of Rock Mechanics and Mining Science, V.62, 86-95p., 2013
- [29] ZILIOTTO, D.M., FERREIRA, M.C.: *“Lidar com gente é muito complicado”: Relações socio profissionais de trabalho e custo humano da atividade em teleatendimento governamental*. Revista brasileira de Saúde Ocupacional, 31(114), 135-148, 2006.

7 Apêndice

Rotina utilizada para obtenção das cadeias das condicionais completas da distribuição *a posteriori*, que utiliza da iteração entre softwares R e Rjags:

```
dados<-read.table("",h=T,dec="")
attach(dados)
n<-nrow(dados)

M<-cbind(x0,x1,x2,...) # Matriz de Covariáveis, no qual x0 é um vetor de "uns"

Y<-dados$y          # Variável resposta

p=ncol(M) # Número de parâmetros incluindo o intercepto (beta0)

require(rjags)      # Pacote necessário para obtenção das cadeias das
                    # condicionais completas a posteriori
data=list(Y=Y,M=M,n=n,p=p)
init=list(beta=c(rep(0,p))) ## chutes iniciais para os parâmetros

modelstring="
  model {
    for (i in 1:n) {
      logit(P[i])<-inprod(M[i,],beta)
      Y[i]~dbern(P[i])
    }
    for (j in 1:p) {
      beta[j]~dnorm(0,1000000)    # Priori
    }
  }
"

model=jags.model(textConnection(modelstring),data=data,init=init)
update(model,n.iter=10000)
```

```
output=coda.samples(model=model,variable.names=c("beta"),n.iter=120000,thin=10)
print(summary(output))
plot(output)

require(coda)
m<-as.matrix(mcmc.list(output)) # Extrai as cadeias do objeto output
```