UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

Ana Carolina Cunha Bueno

Ketlin Hoffmam Padilha

Aplicação de Técnicas de Machine Learning na Predição de Clientes Inadimplentes

CURITIBA
2018

UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

Ana Carolina Cunha Bueno

Ketlin Hoffmam Padilha

Aplicação de Técnicas de Machine Learning na Predição de Clientes Inadimplentes

Trabalho de Conclusão de Curso apresentado à disciplina Laboratório B do Curso de Estatística do Setor de Ciências Exatas da Universidade Federal do Paraná, como exigência parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Eduardo Vargas Ferreira

CURITIBA
2018

RESUMO

Nos dias atuais o Brasil tem vivido uma crise econômica que afeta grande parte da população. Com este cenário, fica cada vez mais difícil entender o perfil da população propensa a se tornar inadimplente, após a concessão de crédito por parte de instituições financeiras. Com isso em mente, o objetivo deste trabalho é criar uma metodologia capaz de prever quais clientes inadimplentes realizarão o pagamento de suas dívidas, e compará-la com algumas abordagens usualmente utilizadas. Os clientes foram analisados durante vinte e quatro meses e identificados como bons ou maus pagadores. O conjunto de dados foi cedido por um banco comercial privado e dispõe de oitocentos e vinte e uma variáveis com informações de aproximadamente um milhão de clientes. A partir deste banco de dados, foi extraída uma amostra aleatória de quarenta e cinco mil clientes e realizadas diversas análises com o intuito de aumentar a discriminação de bons e maus clientes. Os dados foram pré-processados e formatados utilizando algoritmos de *Machine Learning* com o objetivo torná-los mais informativos. Comparamos as abordagens via Regressão logística e *Random Forest*. A avaliação da qualidade das metodologias se deu através da área abaixo da curva (AUC). Como conclusão, pode-se dizer que as técnicas de *Machine Learning* aumentam satisfatoriamente o desempenho das predições.

Palavras-chave: *Machine Learning*, Engenharia de Características, Modelos Preditivos.

Sumário

1 INTRODUÇÃO

O Brasil tem vivido um momento de incertezas econômicas que afetam a grande maioria da população. Os números não deixam dúvidas com relação a crise econômica, a taxa de desemprego no país atingiu em média 12,7% no primeiro semestre de 2018. Segundo o IBGE, o número de desempregados no país ultrapassa 13,2 milhões de pessoas. São vários os possíveis motivos que levaram o Brasil a enfrentar uma crise econômica, como a inflação, desemprego, corrupção e desvalorização da moeda. Em 2018, o número de inadimplentes chegou a 61,8 milhões no primeiro trimestre. Segundo levantamento da Serasa Experian, 40,3% da população adulta possui dívidas em atraso. Os bancos também são afetados pelo crescente número de inadimplentes no país. Expostas a maiores riscos na concessão de crédito, as instituições bancárias tentam limitar a exposição aos maus pagadores, dificultando o acesso ao crédito e elevando os juros. Porém, com o aumento dos juros, menos pessoas são capazes de adquirir empréstimos ou de continuar pagando a dívida já contratada anteriormente.

Uma instituição financeira é uma organização cuja finalidade é fazer o intermédio entre o cliente e algum tipo de serviço do mercado financeiro, tais como empréstimos, investimentos, operações de câmbio, financiamentos, entre outros serviços. Bancos comerciais, corretoras de valores, bancos de investimento são exemplos de instituições financeiras. Os bancos comerciais são instituições públicas ou privadas que têm como principal objetivo proporcionar suprimento de recursos necessários para financiar, a curto e a médio prazos, o comércio, a indústria, as empresas prestadoras de serviços, as pessoas físicas e terceiros. De forma geral, os bancos comerciais captam recursos de investidores (agentes superavitários) e repassam esses recursos aos tomadores de crédito (deficitários). Política de crédito é um conjunto de normas e critérios utilizados pelas empresas para tornar viável o financiamento ou o empréstimo para seus clientes. Pode ser útil para diversas finalidades, tais como aumentar o número de clientes, focando nas necessidades da empresa, ou minimizar os riscos da inadimplência.

A inadimplência trata-se do descumprimento da obrigação financeira, como o não pagamento de bens ou serviços até sua data de vencimento. Do ponto de vista jurídico, é o não cumprimento dos termos do contrato feitos em comum acordo entre as partes. Cobrança equivale a reaver valores perdidos, recuperar, reabilitar crédito. Possui um importante papel no ciclo financeiro das empresas, pois visa melhorar o fluxo de caixa e

minimizar as perdas de negócios futuros. Segundo Leoni e Leoni (1997), “a cobrança é uma função importantíssima em qualquer organização empresarial, pois, afinal, é o retorno do dinheiro ou do capital investido”.

Grande parte das instituições financeiras utilizam técnicas estatísticas, mais especificamente regressão logística para prever a probabilidade de pagamento de um determinado cliente. Entretanto, essa não é a única opção, e nos dias atuais novas técnicas vem ganhando grande destaque na área de modelos preditivos, um exemplo são abordagens advindas do *Machine Learning* (em português, aprendizado de máquina). A utilização das ferramentas de *Machine Learning* proporciona inúmeras vantagens como processar, analisar e prever rapidamente grandes quantidades de dados, com uma precisão, geralmente, superior às abordagens convencionais.

Muitos fatores podem afetar o sucesso de uma análise de dados, como por exemplo, a qualidade dos dados (nível de informação que apresentam), relevância destas, confiabilidade, dentre outros. Todos esses fatores afetam o desempenho de qualquer abordagem, tornando a busca pelo conhecimento através dos dados muito mais difícil e moroso. Por este motivo é necessário desprender um tempo considerável na preparação e aprimoramento dos dados, antes de apresentá-lo aos algoritmos.

2 OBJETIVO

O objetivo do trabalho é apresentar alternativas para melhorar o desempenho dos modelos preditivos através de técnicas modernas de *Machine Learning* e compará-las com abordagens tradicionais. De modo específico, os objetivos são:

- Apresentar uma alternativa para o tratamento de dados faltantes e comparar o desempenho dos algoritmos com e sem esta mudança;

- Apresentar uma alternativa para seleção de variáveis através da Regressão Lasso e compará-la com o método *Foward* (método convencional da estatística);

- Gerar novas variáveis através de componentes principais e indicadores de outliers;

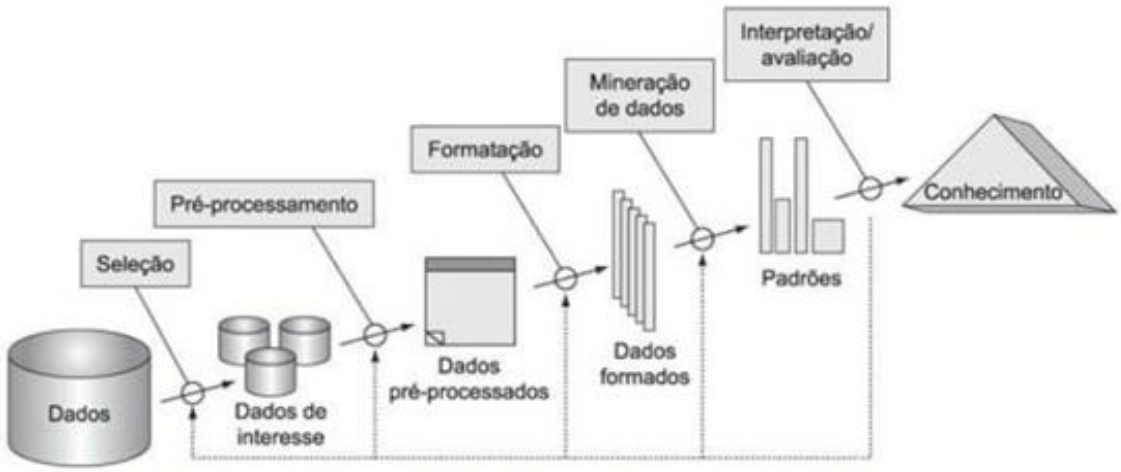
- Por fim, avaliar o desempenho da predição do algoritmo *Random Forest* com a Regressão Logística, considerando diferentes cenários.

3 MATERIAL E MÉTODOS

Neste capítulo apresentaremos algumas informações úteis para o pleno entendimento das análises que realizamos. Trataremos desde o processo de descoberta de conhecimento a partir dos dados, até os modelos utilizados.

3.1 Processo KDD

O processo KDD (*Knowlegde Discovery in Database*) trata-se do percurso que o dado percorre até virar conhecimento. O processo é utilizado para identificar padrões em grandes bases de dados e é dividido em cinco etapas. As etapas são demonstradas na figura 1 e especificadas a seguir.



Fonte: Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic.

Figura 1 – Ilustração das etapas do processo KDD.

1) A **seleção dos dados** é a primeira etapa do processo KDD. Trata-se de uma etapa muito importante, pois é nela que são selecionados os dados para a modelagem. Se os dados selecionados não forem de qualidade, no momento da modelagem podem faltar informações úteis ou gerar informações que não serão utilizadas. A realização incorreta dessa etapa pode impactar no tempo e gerar custos desnecessários. Um modo eficaz para a condução dessa etapa é definir com clareza o objetivo e entender bem os dados para

avaliar sua utilidade. Nessa etapa são definidos os dados de interesse, ou seja, aqueles que serão utilizados em cada momento da análise.

2) A segunda etapa do processo é o **pré-processamento dos dados**, em que o principal objetivo é fornecer dados de qualidade para a realização da modelagem. Caso existam informações inconsistentes pode-se realizar correções, como:

i. Eliminação de dados inconsistentes: algumas variáveis podem apresentar problemas causados por erros humanos, erros na importação de tabelas, etc.

ii. Dados Faltantes: existem alternativas para a correção de dados faltantes. Esses podem ser imputados ou mesmo excluídos. Esse processo é necessário, pois alguns modelos não podem ser executados se a base possuir dados faltantes (trataremos desse assunto com mais detalhes na Seção 3.2).

iii. Selecionar variáveis ou diminuir sua dimensionalidade: esse processo é útil para bancos de dados muito grandes. A seleção de subconjuntos pode ser utilizada para remover informações irrelevantes. Dessa forma, diminuímos a dimensão dos dados, permitindo que os algoritmos de aprendizagem funcionem com maior rapidez e precisão (abordaremos esse assunto na Seção 3.3 e 3.5).

iv. Tratamento de *outliers*: antes de decidir o que deve ser feito com as informações atípicas, comumente denominadas de *outliers*, é conveniente conhecer as causas que levaram seu aparecimento. Em muitos casos, essas anomalias podem ser causadas por erros de medição ou execução, comportamento não convencional de determinadas observações, etc. Após a identificação dos outliers, as informações podem ser eliminadas ou tratadas para que o resultado não apresente efeitos não confiáveis (abordaremos esse assunto com mais profundidade na Seção 3.4).

3) A etapa de **formatação dos dados** é importante pois modifica os dados para que os algoritmos funcionem com qualidade. Alguns exemplos de ações a serem tomadas na etapa de formatação são apresentados a seguir:

i. Transformação de variáveis: por exemplo, comportamento assimétrico positivo, pode-se aplicar a transformação logarítmica de forma que sua distribuição se torne mais simétrica. Tais abordagens tendem a fornecer mais informações ao algoritmo.
ii. Criação de um novo atributo: pode-se criar um atributo a partir de outros. Por exemplo, se quer trabalhar com o índice de massa corporal (IMC) mas as informações coletadas são altura, peso, sexo e idade. A partir de um cálculo envolvendo essas informações, pode ser criado um novo atributo.
iii. Agregação de variáveis: utilizada quando é necessária a modificação da apresentação das variáveis. Por exemplo, a transformação de vendas mensais em diárias.
iv. Discretização de atributos: pode-se transformar um dado contínuo em discreto caso indique uma melhoria da variável com essa alteração.

4) A **mineração dos dados** é o quarto passo do processo KDD no qual os dados se transformam em informações. Essa etapa diz respeito a modelagem onde é possível encontrar padrões nos dados e a partir deles, realizar inferências.

5) A quinta e última etapa do processo é a **interpretação e avaliação** dos resultados obtidos na mineração dos dados. A partir desse passo é possível tirar conclusões e obter conhecimento através dos dados.

3.2 Dados Faltantes

Um problema comum em análises de dados é a ocorrência de dados faltantes, e sua solução não é tão trivial quanto parece. A simples exclusão dos dados faltantes pode gerar inferências equivocadas e perda de informações importantes para análise. Por este motivo, é importante estabelecer estratégias para lidar com esse tipo de dado e utilizar técnicas adequadas para contornar este problema. Existem várias abordagens que envolvem imputação de dados, cujo objetivo é substituir os dados ausentes por

estimativas dos mesmos. Antes de entendê-las, devemos conhecer o mecanismo gerador destes dados faltantes:

a) Estes podem ser completamente aleatórios MCAR (*Missing Completely at Random*) em que a falta dos dados não está ligada as demais variáveis. Por exemplo, quando a captura dos dados para de funcionar por um período de tempo. Esta situação, por exemplo, pode se dar quando uma pessoa participa de uma pesquisa, mas muda de cidade e não pode mais continuar o processo. Neste caso, a exclusão das linhas não faria falta, pois não haveriam informações adicionais na realização da imputação dos dados.

b) Os dados faltantes também podem ser aleatórios MAR (*Missing at Random*) em que a falta dos dados está relacionada a algumas das demais variáveis. Por exemplo, pessoas com alta renda tendem a não querer responder informações sobre seus bens, como número de televisores, casas etc. Neste caso, é possível prever qual seria o valor do dado faltante através das covariáveis.

c) Também existem dados faltantes do tipo MNAR (*Missing Not at Random*) em que o dado é não aleatório e a omissão depende da própria variável cujo dado é omisso. Por exemplo, dados faltantes sobre utilização de drogas por parte dos funcionários pode ser ocasionado porque estes funcionários utilizam drogas, por isso não participam dos exames. Nesse caso, as variáveis observadas não explicam completamente a omissão dos dados, por isso o método de imputação não funciona muito bem.

A imputação de dados faltantes e suas análises são uteis para dados do tipo MAR, pois significa que se pode utilizar a informação das outras variáveis para predizer esse dado faltante. Neste caso, ao realizar uma imputação deve-se supor que os dados faltantes são do tipo MAR.

3.3 Componentes Principais

A análise de componentes principais é uma técnica multivariada, a qual permite condensar as informações presentes em um grande número de variáveis através da criação de algumas variáveis, denominadas componentes, que visam captar o máximo da variabilidade presente nas variáveis originais. Geralmente as primeiras componentes

principais são as mais importantes pois explicam a maior parte da variação total dos dados. Sendo específico, a primeira componente principal de um conjunto de características X_1, X_2,…,X_p é a combinação linear normalizada definida por:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$

Em que ϕ são as cargas da primeira componente principal que maximizam a $Var(Z_1)$.

3.4 Outlier

Um dos problemas mais comuns no pré-processamento dos dados é a ocorrência de *outliers*. Estes são elementos que não obedecem a um padrão típico dos dados a qual eles pertencem. Existem diversas aplicações ao identificar tais anomalias nos dados, como a identificação de fraudes, diagnósticos de falhas, identificação de comportamento do consumidor, entre outros. A preocupação com essas observações atípicas é antiga e a sugestão inicial era eliminar os dados da análise. Esse procedimento ainda é muito utilizado mesmo existindo outras formas de lidar com esse problema. Os *outliers* podem conter informações de extrema importância em relação aos dados e ao excluí-los, podemos obter um resultado distorcido da realidade. Da mesma forma que ao mantê-los na base, um efeito desproporcional pode ser causado sobre os resultados e repercutir interpretações equivocadas. Neste trabalho, utilizaremos como detecção de *outliers* o método de agrupamento conhecido como *Clustering*, em que o algoritmo de aprendizado analisa os dados fornecidos e os agrupa em clusters segundo algum critério de similaridade. Este método é conhecido como aprendizagem não-supervisionada.

Existem diversos algoritmos de agrupamento de dados para realizar a aprendizagem não-supervisionada. Dentre eles, podemos citar: K-MEANS [Jain, 2010], CLUSTER HIERÁRQUICO [Jain et al., 1999; Sander et al., 2003], DBSCAN [Ester et al, 1996], entre outros. Os métodos K-means e cluster hierárquico são adequados para clusters compactos (de formato esférico ou convexo). Como os dados do presente estudo refletem a vida real e podem conter clusters de formato arbitrário, de diferentes tamanhos e conter ruído, utilizaremos o algoritmo DBSCAN (*Density Based Spatial Clustering of Application with Noise*). O método agrupa pontos que são próximos uns dos outros com base em uma medida de distância (geralmente euclidiana) e um número mínimo de

pontos. Também classifica como *outliers* os pontos que estão em regiões de menor densidade. O algoritmo DBSCAN basicamente requer dois parâmetros de entrada: *eps* (menor distância entre dois pontos) e *minPoints* (número mínimo de pontos para formar uma região densa). Estes parâmetros são utilizados no KNN (*K nearest neighboors*) que possui como foco principal reconhecer padrões. O centro de seu funcionamento está em descobrir o vizinho mais próximo de uma dada instância.

3.5 Regressão Lasso

Ao nos deparamos com um grande número de variáveis preditoras que explicam determinada resposta, são necessários mecanismos de escolha das melhores variáveis. Por exemplo, no caso de regressão linear, são estimados os parâmetros β, e as variáveis associadas às estimativas diferentes de zero que entrarão no modelo. Sob certas condições, os estimadores de Mínimos Quadrados Ordinários (MQO) são não viesados e de variância mínima. Ou seja, dentro da classe dos estimadores não viciados, procura-se aquele de menor variabilidade. Em *Machine Learning* o interesse está na predição de novas observações (e não muito na interpretação dos parâmetros). Por isso, pode-se permitir certo viés nas estimativas dos parâmetros, a fim de obter consideráveis decréscimos na função que se deseja minimizar, geralmente denominada de função custo, que no nosso caso será dada pela perda quadrática:

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

A parcela que acompanha o parâmetro λ representa a função de penalização, cujo papel é manter as estimativas dos betas próximas de zero (regulando-as). Assim, pode-se dizer que ao usar a abordagem Lasso, de alguma forma, encurta-se o valor que o beta receberia dentro de uma regressão linear convencional, ou seja, ele está recebendo um vício nas estimativas originais, e por isso não é prudente realizar qualquer interpretação dos parâmetros. A vantagem em perder a interpretação dos parâmetros pode ser compensada pelo poder preditivo que se adquire.

Na regressão regularizada Lasso, o λ é o primeiro parâmetro a ser estimado. Logo depois, utiliza-se esse λ como valor fixo durante o processo de estimar os β's do modelo. Dependendo do valor do λ selecionado, algumas estimativas dos parâmetros podem ser

zero, assim se pode dizer que o Lasso funciona como um selecionador de variáveis. A estimativa do melhor modelo e do λ é feita via validação cruzada. Este método consiste em dividir os dados em *k* partes iguais, ajusta-se o modelo utilizando *k−1* partes, e a parcela restante fica destinada à validação. Esse processo é repetido *k* vezes (em cada momento uma partição diferente será a validação), e em seguida os resultados são combinados obtendo a média dos erros obtidos. O λ e o modelo escolhido será aquele que apresentar menor erro de validação. A figura 2 exemplifica o processo da validação cruzada k-fold.

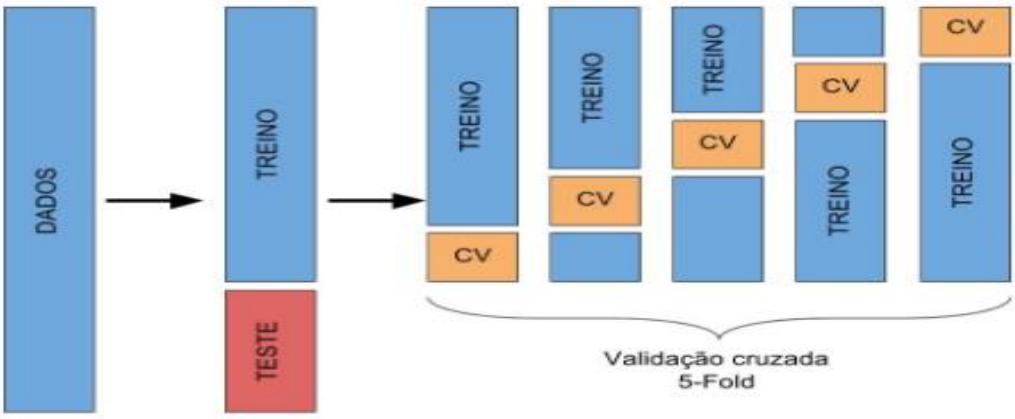


Figura 2 – Esquema representativo da validação cruzada 5-fold.

Para encontrar o menor erro, deve-se indicar ao algoritmo qual a melhor medida para avaliar os erros. Existem diversos tipos de avaliação de erros que podem ser escolhidos como a acurácia, *deviance*, entre outros. É necessário possuir entendimento de quais são as métricas e quando devem ser utilizadas. Neste trabalho utilizaremos como critério a curva ROC.

3.6 Curva ROC

A curva ROC (*Receiver Operating Characteristic*) é um método gráfico que mede a capacidade de classificar corretamente um dado com característica dicotômica (por exemplo um cliente bom ou mau pagador). A validade de uma boa predição está na capacidade de identificar o maior número possível de acertos (resultados positivos verdadeiros) e minimizar os erros (falsos resultados positivos). Em outras palavras, maximizando a sensibilidade e a especificidade e minimizando os falsos acertos. A curva ROC avalia esses resultados plotando todos os valores de acertos verdadeiros

(sensibilidade) no eixo y, em relação a proporção de falsos acertos (1-especificidade) no eixo x. A partir do resultado da curva ROC, escolhe-se o ponto de corte referente a combinação da sensibilidade e 1-especificidade que mais se aproxima do canto superior esquerdo do gráfico. Quanto maior a área abaixo da curva, melhor a capacidade de classificação. Veja um exemplo na Figura 3.

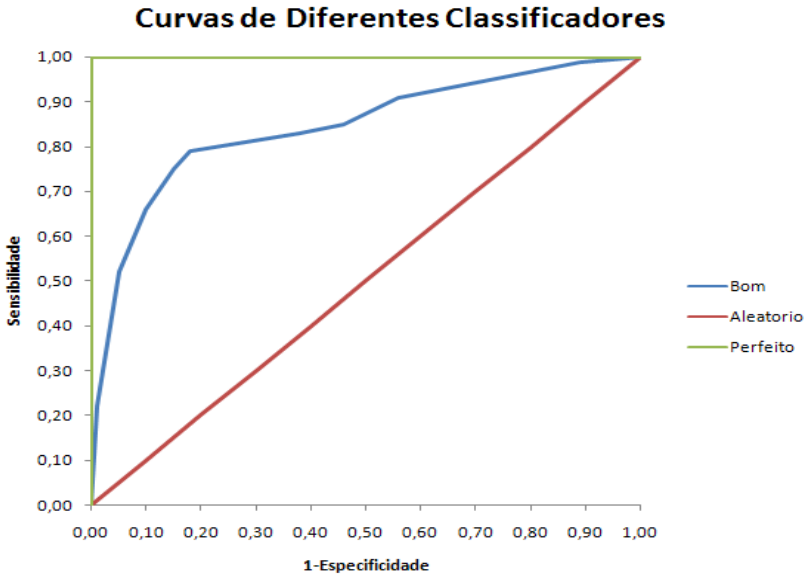


Figura 3 – Exemplo simulado de uma curva ROC.

3.7 Random Forest

Random Forest é um algoritmo de aprendizagem que produz excelentes resultados e devido a sua simplicidade é muito utilizado em predição de respostas contínuas e discretas. O modelo que tem como base o princípio da árvore, criando-se b árvores diferentes a partir dos dados fornecidos, e esse b é um parâmetro definido pelo pesquisador. Cada uma das b árvores é criada utilizando m variáveis selecionadas aleatoriamente através do *bootstrap* (de p disponíveis), ou seja, não são utilizadas todas as variáveis da base para criar as árvores. O valor usualmente utilizado para m é a raiz quadrada do número de variáveis existentes nos dados completos. A figura 4 apresentada a seguir representa este processo. O fato de criar b árvores diminui a variabilidade da predição, e por não utilizarmos todas as variáveis em sua construção, descorrelacionamos as árvores. Para mais detalhes veja James et. al. (2013).

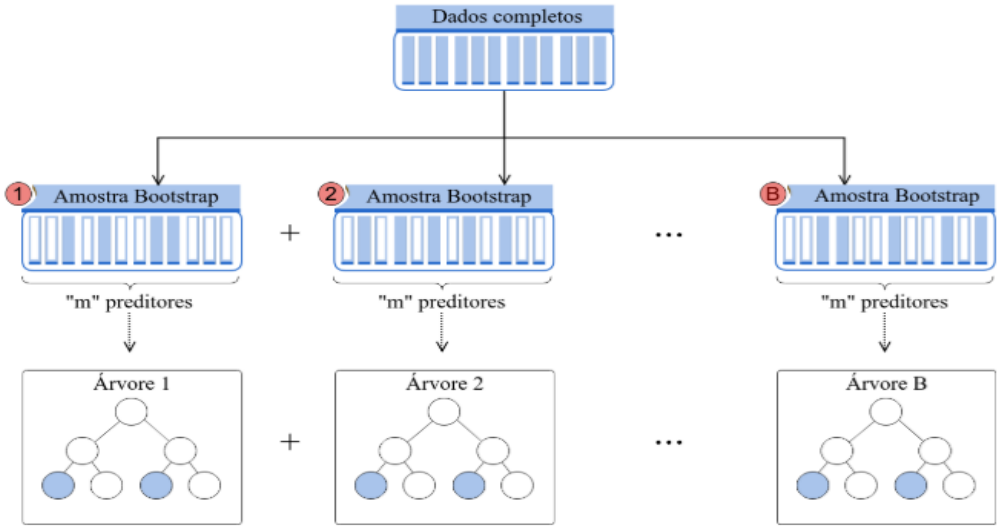


Figura 4 – Fluxograma representativo do *Random Forest*.

3.8 Regressão Logística

A regressão logística é um dos principais modelos utilizados quando a variável resposta é binária ou dicotômica. O modelo tem como objetivo relacionar uma variável resposta com um conjunto de variáveis explicativas. O modelo de regressão logístico utiliza a função de ligação logito em um modelo linear generalizado binomial. O modelo é dado pela seguinte equação:

$$P(Y = 1 \mid X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Ou seja, calculamos a probabilidade de Y ser igual a 1 dado um conjunto de variáveis X. O logaritmo da razão entre os termos P(Y=1|X) e 1 - P(Y=1|X) fornece um modelo linear:

$$log\left[\frac{P(Y=1 \mid X)}{1-P(Y=1 \mid X)}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

3.9 Seleção de Variáveis

Na maioria das análises utilizando dados em grandes dimensões é necessário determinar um subconjunto de variáveis independentes que expliquem de melhor forma a variável resposta. Para isso, um algoritmo de seleção de variáveis deve ser utilizado. Para adicionar ou remover variáveis, o critério é geralmente baseado na estatística F, comparando modelos com e sem as variáveis em questão. Existem três procedimentos automáticos (não somente estes): (i) método *forward*, (ii) método b*ackward* e (iii) método

s*tepwise*. Neste trabalho, o método utilizado na seleção do modelo logístico será o *forward* que supõe que não existem variáveis no modelo, apenas o intercepto, e a cada iteração adiciona-se uma variável. Outra abordagem utilizada é via Regressão Lasso, pois como mencionado funciona também como selecionador de variáveis.

3.10 SMOTE

Lidar com distribuição desbalanceada nas amostras é um problema para realizar o reconhecimento de uma determinada classe. Isso ocorre quando o número de exemplos que determinam uma classe é muito menor que o outro. O método denominado SMOTE (*Synthetic Minority Over-sampling Technique*) realiza superamostragem da classe minoritária criando novas amostras balanceadas. Esta proposta melhora a representatividade das classes e altera o comportamento do algoritmo de tal forma que reduz o risco de classificar incorreta a classe com menor frequência.

3.11 Conjunto de Dados

O conjunto de dados utilizado no presente trabalho foi disponibilizado por um banco comercial privado, que dá acesso a grande parte da população brasileira a créditos como: financiamento de automóveis, crédito imobiliário, cartão de crédito, conta-corrente, entre outros. Os dados apresentam 821 características (variáveis) de um milhão sessenta e nove mil e duzentos e setenta clientes inadimplentes (com mais de sessenta dias em atraso), que descumpriram com a obrigação de pagamento na data do vencimento. Os clientes da base em questão foram monitorados durante vinte e quatro meses após o atingimento da faixa de sessenta dias em atraso. O objetivo foi verificar se houve pagamento do produto em atraso em algum momento durante o período de observação. Para as análises deste trabalho utilizamos uma amostra aleatória de quarenta e cinco mil observações, separada em treinamento (para treinar os algoritmos) e validação (para avaliar a qualidade do treinamento).

3.12 Recursos Computacionais

Utilizamos o *software* R, versão 3.5.1 para ajustar os modelos aos dados descritos por meio de pacotes apropriados como: CARET (para os algoritmos de *Machine Learning)*, MICE (para imputação), DBSCAN (para detectar outliers), entre outros.

4 RESULTADOS E DISCUSSÃO

4.1 Pré-Processamento e Formatação dos Dados

O pré-processamento não necessariamente é feito antes da formatação dos dados. A separação dentro da teoria acontece devido aos objetivos serem diferentes. O processamento foca em fornecer dados com melhor qualidade e a formatação em torná-los bons para o algoritmo. Sendo assim são apresentadas a seguir as ações de pré-processamento e formatação dos dados dentro de único tópico, localizando-se através dos subtópicos dos assuntos.

Realizamos a seleção de variáveis e redução de dimensionalidade: como muitas variáveis do banco de dados possuíam uma grande concentração de dados faltantes, o primeiro passo foi eliminar variáveis que possuíam mais de 30% dos dados *missing*, uma vez que qualquer procedimento diferente poderia piorar o resultado final por adicional excesso de ruído. Após este passo, o conjunto de dados passou a ter trezentas e cinquenta e duas variáveis. Excluímos também variáveis que possuíam pouca variabilidade, pois se as variáveis não variam, consequentemente não trazem informação suficiente para a resposta. Por exemplo, a variável **quantidade de cheques sem fundo devolvidos na segunda apresentação (V712)** que foi eliminada é mostrada no gráfico da figura 5. A **variável resposta (V2)** apresenta o perfil dos clientes, em que 0 trata-se de clientes maus pagadores e 1 bons pagadores.

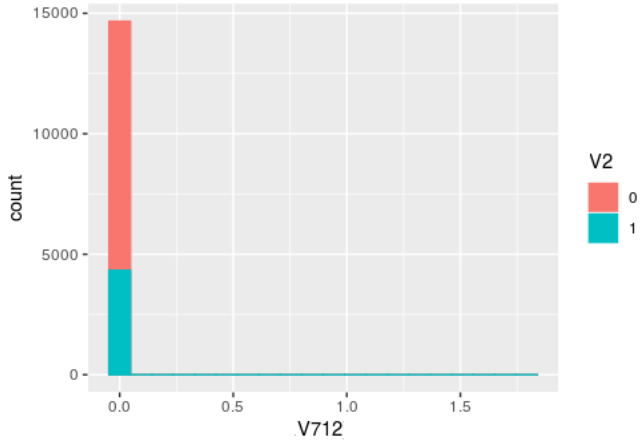


Figura 5 – Distribuição da variável V712

Ao nos deparamos com variáveis muito semelhantes optamos manter na base apenas uma delas. Como exemplo as variáveis: **percentual de contratos renegociados**

nos últimos 6 meses (V622) e **percentual de contratos renegociados nos últimos 12 meses (V634)**. Como as duas variáveis não possuem diferenças significativas, a variável escolhida para permanecer na base foi a V622 por apresentar dados mais atualizados.

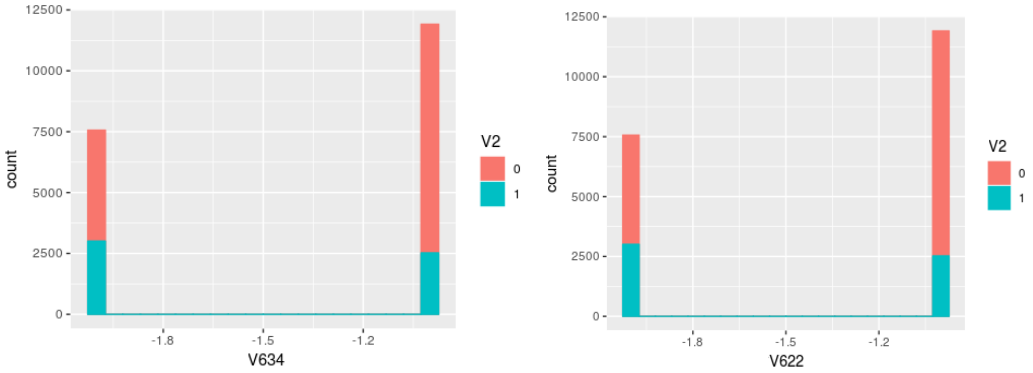


Figura 6 – Distribuição da variável V634 e V622.

Além disso, foram transformadas variáveis aplicando *log*. Como exemplo, é apresentada na figura 7 a variável **quantidade de cheques devolvidos (V716).** Com a transformação, espera-se que os resultados sejam mais informativos.

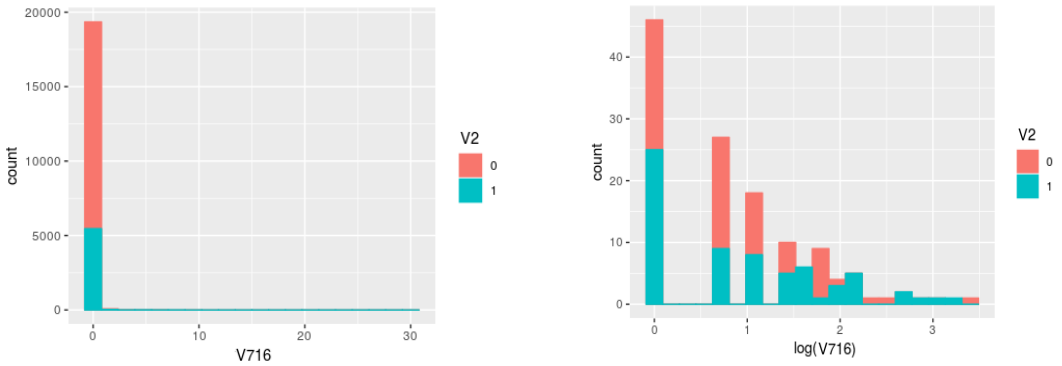
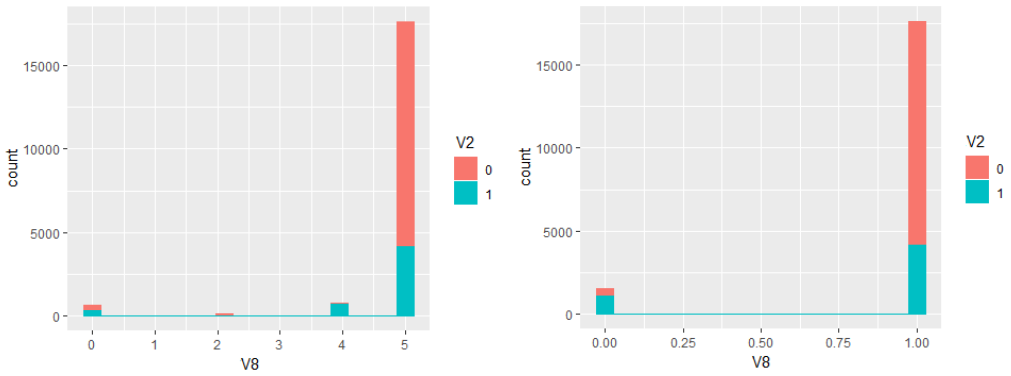


Figura 7 – Distribuição da variável V716 antes e depois da aplicação do log.

Discretizamos variáveis, de forma a apresentaram melhores separações entre os grupos de bons e maus pagadores. Na figura 8(a) é possível verificar que a variável **pior grau do restritivo ativo (V8)** original possuía seis resultados e grande concentração em um campo. Foram agrupados os valores de 0 a 4 e a nova variável é apresentada na figura 8(b). Podemos observar que após a modificação houve melhor distinção da variável resposta nas categorias 0 e 1.



(a) (b)

Figura 8 – Distribuição da variável V8 antes e após a discretização.

Restaram 65 variáveis nas quais foram realizadas 21 discretizações e 13 aplicações de *log*. As variáveis restantes permaneceram com sua forma original.

4.2 Dados Faltantes

O gráfico da figura 9 (à direita) apresenta a frequência de dados faltantes das variáveis. Ao observar o gráfico verificou-se que as variáveis **tempo de relacionamento (em meses) da abertura da conta (V810)**, **identificação de utilização do internet banking (V655), identificação de utilização de caixa automático (V680)** e **pior grau de restritivo ativo, baixado ou decursado (V11)** possuíam dados faltantes. Este gráfico apresenta a frequência de *missing* existente em cada variável. O gráfico da direita apresenta as diferentes combinações das variáveis em termos de dados observados (em azul) e faltantes (em vermelho). Analisando a proporção no eixo das ordenadas do lado direito, é possível observar que a maior frequência das variáveis combinadas não apresenta dados faltantes.

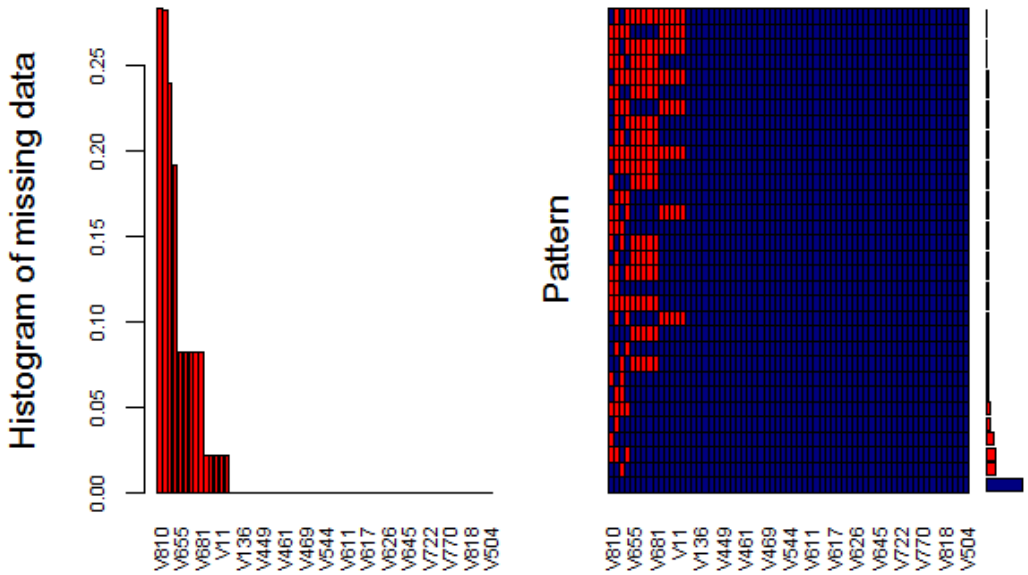
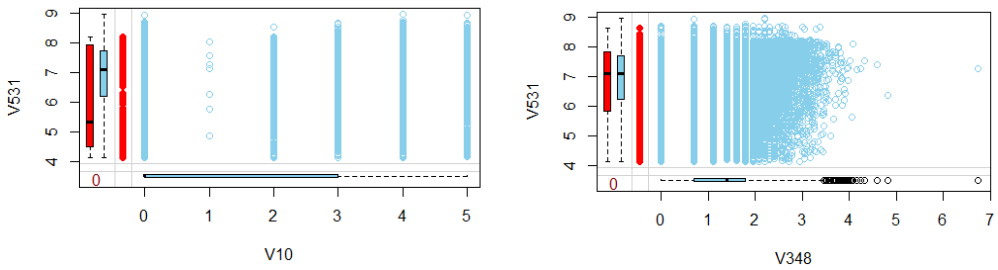


Figura 9 – Frequência dos dados faltantes nas variáveis.

Observando o gráfico da figura 10 podemos verificar se há perda de informação ao excluir os dados faltantes de uma variável. Dois exemplos são apresentados para contextualizar. No eixo Y da figura 10(a) são apresentados dois box plots, o vermelho diz respeito a distribuição da variável **atraso inicial (V531)** na presença de dados faltantes na variável **pior grau do restritivo decursado (V10),** já o azul demonstra o comportamento da V531 quando não existem dados faltantes na V10. Ou seja, o comportamento da V531 é diferente em termos da variabilidade e da mediana entre os box plots. Sendo assim, é possível verificar que ao excluir dados faltantes da V10, alguma informação da V531 também será eliminada. Por outro lado, o gráfico da figura 10(b) mostra que eliminar dados da variável **quantidade de restritivos (V348)** não há impacto na informação da V531. Isso pode ser verificado através da similaridade entre os box plots.



(a) (b)

Figura 10 – Distribuição dos dados com e sem dados faltantes.

Outra análise importante diz respeito à correlação entre as variáveis, como apresentado na figura 11. A cor azul representa correlação positiva entre as variáveis e a cor vermelha, correlação negativa. Por exemplo, observou-se que a variável **percentual máximo de baixa de restritivos (V135)** possui forte correlação com a variável **percentual de baixa de restritivos em 9 meses (V141)**. Consequentemente, caso existam dados faltantes na V135, a V141 possivelmente auxiliará na imputação dessa observação.

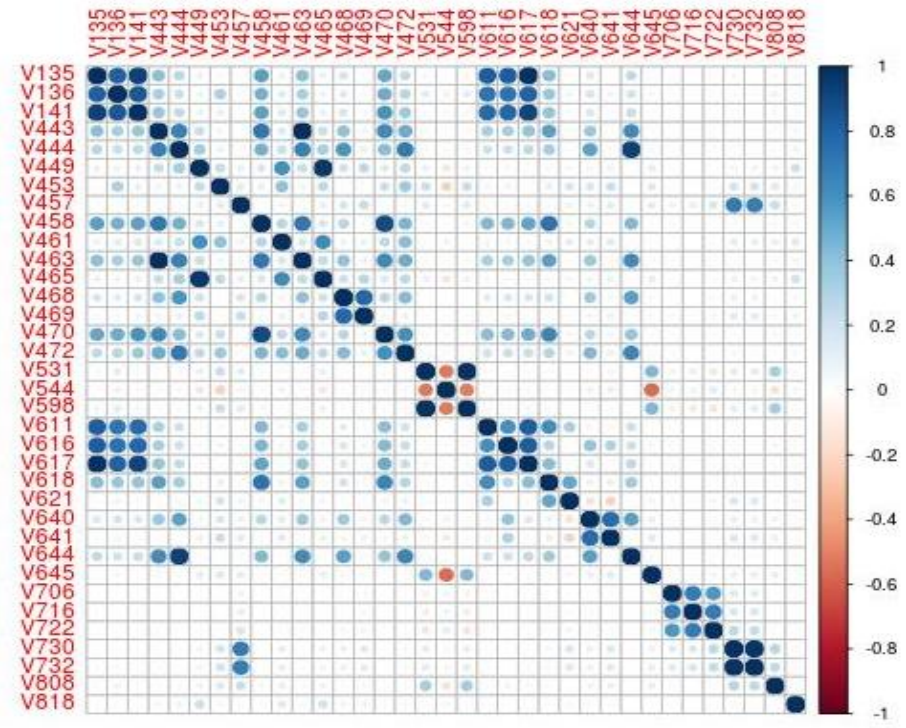


Figura 11 – Representação gráfica da correlação entre as variáveis.

Para proceder o processo de imputação, é sabido que diferentes abordagens podem conduzir a diferentes resultados. Pensando nisso, realizamos a imputação utilizando *Random Forest* e Regressão e os resultados foram divergentes, veja, por exemplo, a variável **quantidade de operações vencidas (V324),** na figura 12. Considerando a não linearidade das distribuições das características (e após um estudo simulado) decidimos por utilizar o *Random Forest* para esta tarefa. A figura 13 apresenta esse resultado, em que a curva em azul refere-se aos dados observados e a vermelha aos dados imputados pelo algoritmo

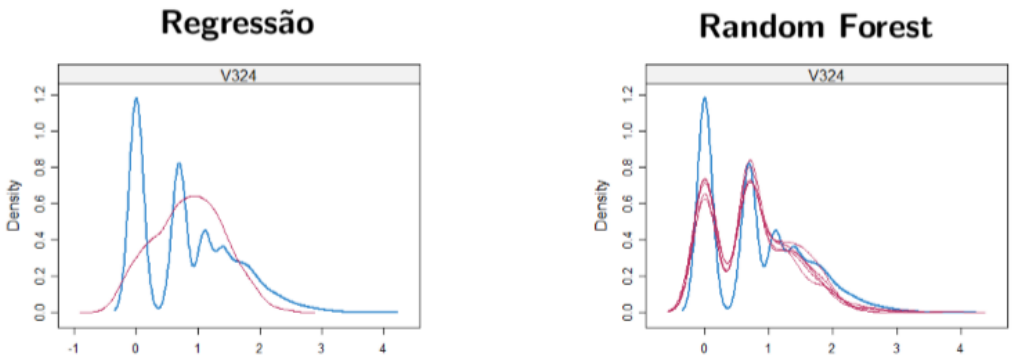


Figura 12 – Imputação da variável V324 com Regressão e *Random Forest*.

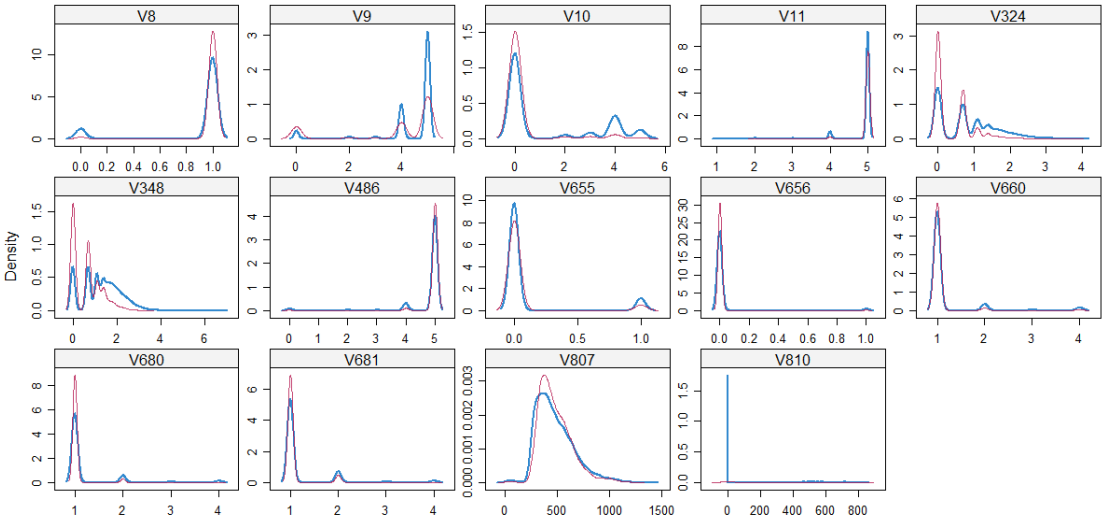


Figura 13 – Imputação de todas as variáveis faltantes usando o método *Random Forest*.

4.3 Regressão Lasso

A regressão lasso foi utilizada para selecionar variáveis com bom poder de predição em meio as 64 variáveis restantes após os processos apresentados anteriormente. O resultado das estimativas do parâmetro de regularização, λ, é apresentado na figura 14. A linha tracejada à esquerda representa os resultados associados ao λ que fornece o erro mínimo de validação cruzada. Já a linha tracejada à direita representa os resultados associados ao λ que fornece o erro que está dentro de um erro padrão.

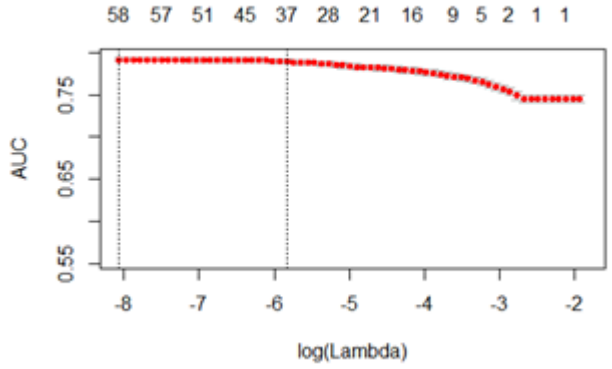


Figura 14 – Gráfico informativo das estimativas dos lambdas.

Optamos pelo λ da direita pois realiza a exclusão de mais variáveis. Portanto, 37 variáveis permaneceram no modelo. Na figura 15 podemos observar o "encurtamento" que acontece nas estimativas dos parâmetros do modelo, conforme o λ muda.

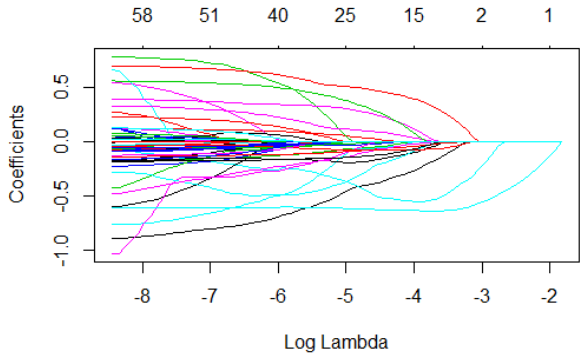


Figura 15 – Variação dos betas do modelo conforme o λ se altera.

A partir da saída apresentada abaixo, pode-se verificar as 37 variáveis que permaneceram no modelo para serem avaliadas. As variáveis que tiveram 0 atribuído foram eliminadas.

```
          V8           V9          V10          V11         V135         V136
-6.616295e-01  8.934660e-02 -4.904886e-02  0.000000e+00 -2.439436e-01  0.000000e+00
        V141         V324         V348         V443         V444         V449
 0.000000e+00  0.000000e+00 -1.230471e-01 -4.643142e-02 -4.736682e-01 -2.437119e-01
        V453         V457         V458         V461         V463         V465
 0.000000e+00  1.152174e-01 -5.008378e-02  0.000000e+00  0.000000e+00  0.000000e+00
        V468         V469         V470         V472         V486         V531
 0.000000e+00  4.941490e-02  0.000000e+00  4.972556e-01 -3.646302e-02 -6.097267e-01
        V544         V554         V598         V602         V611         V612
 2.062707e-01 -1.637893e-01 -1.138730e-01 -1.129039e-03 -9.691687e-03 -5.992714e-01
        V613         V616         V617         V618         V621         V622
-1.065855e-01  0.000000e+00  0.000000e+00 -5.426040e-02 -4.790208e-02  0.000000e+00
        V626         V640         V641         V644         V645         V655
 0.000000e+00  0.000000e+00  3.389165e-01  0.000000e+00 -1.652145e-01  5.994217e-01
        V656         V660         V680         V681         V706         V716
 4.849937e-01  0.000000e+00  0.000000e+00 -5.735403e-02  0.000000e+00 -2.386464e-02
        V720         V722         V730         V732         V764         V770
 0.000000e+00 -2.675483e-02  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
        V791         V807         V808         V809         V810         V818
 0.000000e+00 -6.527292e-04 -3.310992e-02  4.727212e-03 -1.350848e-05  0.000000e+00
        V820         V490         V501         V504
```

Se o valor 0 é atribuído ao β podemos concluir que a variável não possui importância para o modelo. Quanto maior o valor atribuído ao β, maior a significância da variável.

4.4 Análise de Componentes Principais

Por estarmos diante de um grande número de variáveis explicativas, a análise de componentes principais foi utilizada visando produzir uma representação dos dados em dimensões menores sem perda significativa de informação. As variáveis foram padronizadas para que não fossem privilegiadas determinadas características por estarem em outra escala. Foram geradas 35 componentes principais, das quais foram escolhidas para permanecer na análise a CP1, CP2, CP3 e CP4. A escolha foi feita a partir do gráfico abaixo que mostra o quanto de variância as quatro primeiras componentes principais dominam.

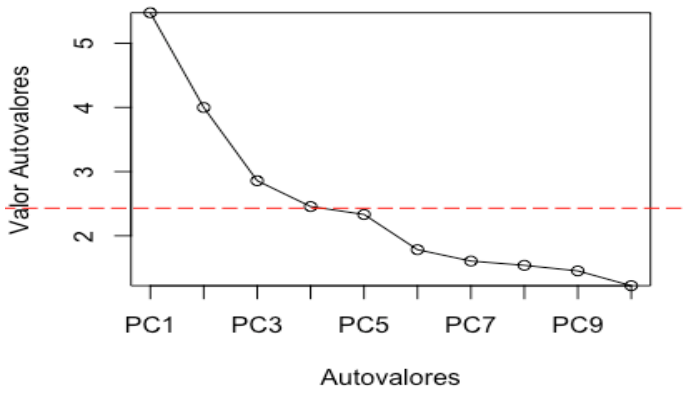


Figura 16 – Variância das componentes principais.

4.5 Análise de Outliers

Com o objetivo de identificar os *outliers* presentes nas observações pré-selecionadas para a aplicação do modelo, utilizamos o algoritmo DBSCAN. Para esse método precisamos de dois parâmetros de tuning: eps e MinPts. Para a escolha ótima dos parâmetros foram calculadas as distâncias via KNN, com K = MinPts, onde o valor eps = 0.2 foi escolhido. O processo de formação de *clusters* foi realizado a partir das componentes principais calculadas anteriormente com os parâmetros de *tuning* definidos. O resultado obtido é apresentado no gráfico da figura 17. Ao todo, foram encontrados 193 clientes com perfil atípico dentre as 25 mil observações selecionadas.

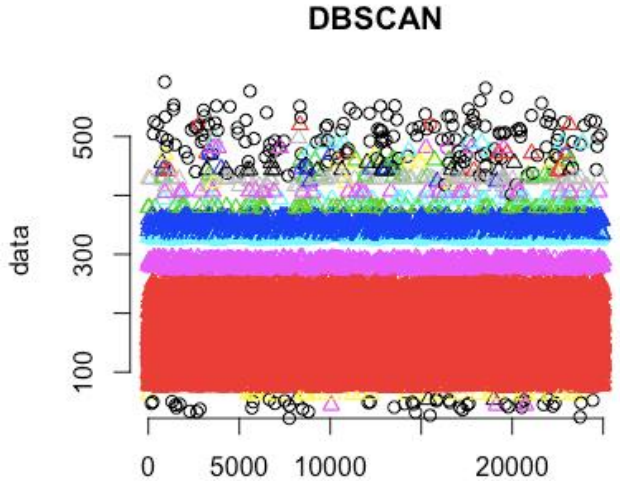


Figura 17 - Formação de agrupamentos identificando os outliers.

O gráfico acima apresenta regiões densas no espaço de dados que são denominadas clusters. Os pontos destacados em preto são considerados outliers.

Para a análise do modelo foi criada uma variável com uma marcação de *outlier* em que 0 representa um cliente sem este perfil e 1 cliente com o perfil.

4.6 Comparação entre Modelos

Para encontrar a melhor abordagem para classificar um cliente em bom a mau pagador, ajustamos os modelos logísticos e *Random Forest* com e sem pré-processamento (tratamento de *missing*, geração de variável *outlier* e componente principal) nos dados. Para os dados sem tratamento de *missing*, a técnica de seleção de dados utilizada foi a *Foward* e para os dados tratados a seleção foi via *Lasso*. Separamos dessa forma a fim de comparar as técnicas utilizadas e seus ganhos em predição. Foram

utilizados vinte e cinco mil dados para treino e vinte mil dados para validação. Como critério de avaliação utilizamos a área abaixo da curva ROC. A tabela abaixo apresenta os resultados obtidos.

Tabela 1: Resultados obtidos através dos métodos.

Índice Localizador	**Métodos**	**Resultado AUC**
1	Regressão logística sem tratamento de *missing*, geração de variável outlier e componente principal. Seleção de variáveis via *Foward.*	0,7413
2	Regressão logística com tratamento de *missing*, geração de variável outlier e componente principal. Seleção de variáveis via Lasso.	0,7598
3	*Random forest* sem tratamento de *missing*, geração de variável outlier e componente principal. Seleção de variáveis via *Foward.*	0,7553
4	*Random forest* com tratamento de *missing*, geração de variável outlier e componente principal. Seleção de variáveis via *Lasso.*	0,7755

Através dos resultados pode-se verificar que o ganho médio referente ao tratamento de *missing*, geração das componentes principais e uso do selecionador de variável lasso foi de 2,0%. Este ganho não se deu pela criação da variável com marcação de clientes com perfil *outlier*, pois a importância desta variável foi muito baixa (ver figura 18). Isto pode ser explicado pelo baixo volume de clientes encontrados com tal perfil (apenas 193) em relação ao tamanho da amostra de treinamento. O modelo que apresentou melhor assertividade ao identificar bons e maus clientes foi o *Random forest* com dados tratados (linha 4). Ao comparar o modelo logístico (linha 2) com *Random forest* (linha 4) pode-se verificar que o ganho na área abaixo da curva foi de 1,57%.

A base de dados possui 77% dos clientes com perfil mau pagador e 23% com perfil bom pagador. Com o intuito de aumentar a qualidade da predição, foi aplicada a técnica SMOTE. A proporção de clientes bons pagadores foi aumentada para 30% e depois para 50%, porém não houve aumento na área abaixo da curva ROC. No entanto observamos que o índice de acerto de bons pagadores (especificidade) aumenta conforme aumentamos a proporção dos bons pagadores na base.

O gráfico da figura 18 apresenta os resultados da importância das variáveis na predição dos clientes inadimplentes. A variável **atraso inicial (V531)** se mostrou a mais importante, seguida das variáveis **máximo do atraso do cliente (V598), componente**

principal 1 (PC1), componente principal 2 (PC2) e componente principal 3 (PC3). Essa ordenação se deu através das variáveis mais importantes em cada árvore criada. Conforme mencionado anteriormente, através do gráfico é possível verificar que a variável com marcação de perfil de clientes *outliers* ficou entre as menos importantes para predição de bons e maus clientes.

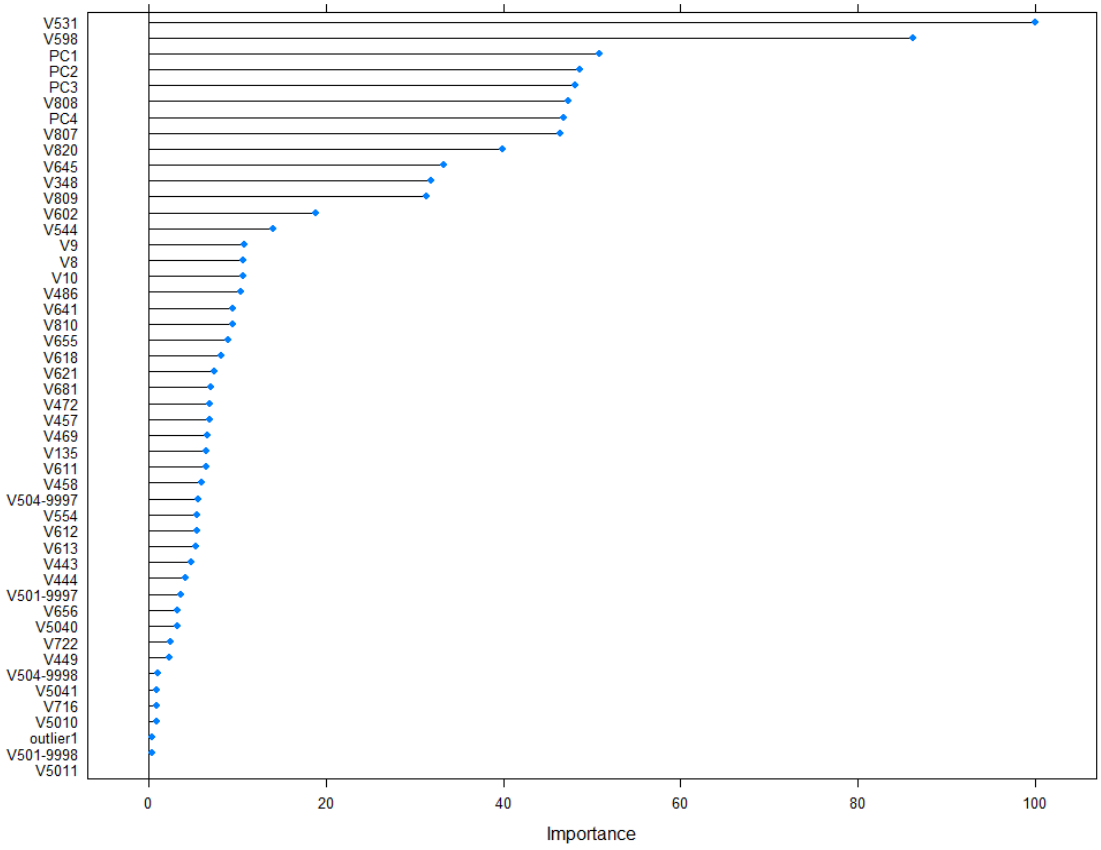


Figura 18 – Gráfico de importância das variáveis.

5 CONSIDERAÇÕES FINAIS

A utilização de técnicas estatísticas com o objetivo de prever o perfil do cliente traz ganhos consideráveis para o setor bancário. Esse trabalho teve como objetivo agregar técnicas de *Machine Learning* que contribuíram significativamente na predição do perfil do cliente inadimplente. A aplicação do processo KDD foi de extrema importância para o trabalho, pois através dele foi possível transformar os dados em conhecimento e promover ganhos importantes. O banco de dados possuía mais de um milhão de observações e um grande número de variáveis. Foi necessário desprender um tempo considerável na preparação e aprimoramento dos dados, antes de apresentá-los aos algoritmos. O tratamento prévio das variáveis, a imputação dos *missings* e a criação de

componentes principais foram essenciais para que os resultados fossem satisfatórios. A identificação de *outliers* não surtiu efeito na predição e isso pode ser devido ao critério de determinação destas observações atípicas. Através dos resultados foi possível concluir que eliminar os dados faltantes gera perda de informação e que a imputação dos dados *missing* colabora para a predição mais assertiva. Também pode-se concluir que o algoritmo *Random forest* obtém melhores resultados quando comparado ao modelo logístico. A maioria dos modelos de negócio capturam apenas uma fração do valor potencial dos dados e existe uma grande variedade de métodos que podem ser utilizados para aumentar o ganho de informação. O trabalho demonstrou que o uso de algoritmos de *Machine learning* oferecem vantagens competitivas em relação as técnicas tradicionais e aumentam a eficiência dos resultados garantindo melhores decisões de negócio. É importante constatar que esse trabalho é apenas o início, e outras questões devem ser aprimoradas a fim de aumentar as predições corretas. Como trabalho futuro sugerimos outras abordagens como métodos *Boosting*, por exemplo, AdaBoosting e outras metodologias de pré-processamento e seleção de variáveis.

REFERÊNCIAS

[1] Introdução à Ciência de Dados: mineração de dados e big data James, G., Witten, D.,

[2] Hastie, T. e Tibshirani, An Introduction to Statistical Learning, 2013

[3] Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. ACM Computing Surveys, 31(3):264–323

[4] K-MEANS A.K. Jain / Pattern Recognition Letters 31 (2010)

[5] Ester M., Kriegel H.-P., Sander J., Xu X.: “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, 1996, pages 226-231.

[6] Sander, J., Qin, X., Lu, Z., Niu, N., and Kovarsky, A. (2003). Automatic extraction of clusters from hierarchical clustering representations. In PAKDD - Pacific-Asia Knowledge Discovery and Data Mining, volume 2637 of LNAI, pages 75–87. SpringerVerlag.