Análise de Credit Scoring

Predição e significância de covariáveis

Prof. Dr. Osvaldo Anacleto Junior

Carolina Alencar
Giovanna Zolin Pinheiro Hayasida
Lucas Yudi Sugi

Introdução

Atualmente é muito difícil sobreviver ao mercado sem utilizar nenhuma tecnologia como as redes sociais, marketing digital, sistemas de gerenciamento, sistemas de banco de dados e etc. Nesse contexto, nota-se que a competitividade entre as empresas está crescendo cada vez mais e aquelas que decidem investir em tecnologia acabam dando um passo à frente da concorrência.

Assim, existem empresas que percebem como a explosão atual dos dados pode ser benéfica ao gerar insights melhorando sua colocação no mercado.

No contexto dos bancos, aqueles que sabem dar crédito a pessoas adimplentes e negá-lo aos inadimplentes com um alto grau de certeza, irão certamente obter lucros e diminuir os prejuízos, respectivamente.

Objetivos

Nosso objetivo é gerar o melhor modelo (classificador) possível para realizar previsão de default (calote) ou não default (não calote) de modo a otimizar a relação lucro/prejuízo do banco.

Para o nosso problema em específico iremos priorizar a predição de default dado que a pessoa é default. Além disso, será avaliado o efeito das covariáveis do modelo na variável resposta.

Considerações iniciais

Este relatório visa ser um complemento aos códigos anexados em conjunto, contendo uma abstração sobre toda a etapa de criação do modelo. Para detalhes mais técnicos ou gráficos, deve-se consultar os códigos fonte que estão bem documentados.

Ademais, salientamos que os programas das análises estão divididos em 4 arquivos:

- AED.py (Análise Exploratória dos Dados em python): Contem a maior parte dos gráficos e análises descritivas.
- Modelagem.r: Criação dos modelos com métodos de seleção de variáveis e avaliação via resíduos.
- Classificacao.r: Teste dos modelos quanto a classificação.
- Interpretacao.r: Interpretação das covariáveis do modelo.
- Utils.r: Funções úteis que foram utilizadas durante todo o processo.

Também apresentamos os comandos das bibliotecas necessárias em R que são necessários para executar todos os códigos:

- install.packages('MASS')
- install.packages('car')
- install.packages('statmod')
- install.packages('reshape2')
- install.packages('caret')
- install.packages('ROCR')
- install.packages('ggplot2')
- install.packages('readxl')
- install.packages('glmnet')
- install.packages('pROC')

Desenvolvimento

O desenvolvimento do projeto foi executado em etapas de maneira iterativa, sendo pensado da seguinte maneira:

- Análise exploratória: Entender bem o conjunto de dados assim como resolver possíveis problemas.
- Modelagem: Criar vários modelos preditivos utilizando inicialmente técnicas de seleção de variáveis. Após essa etapa, adicionar algumas iterações para ver se há melhoras no modelo. Por fim, avaliá-los quanto a colinearidade e resíduos.
- Classificação: Utilizar métricas como AUC para verificar qual modelo será o melhor classificador.
- Interpretação: Realizar interpretação das variáveis dos melhores modelos preditivos.

Análise Exploratória

Antes de construir um classificador devemos avaliar o conjunto de dados disponíveis pois a sua qualidade irá afetar nossos modelos preditores. Assim, inicialmente foi verificado se na base de dados havia problemas como:

- Valores faltantes.
- Valores impossíveis como idade 200, renda -1000, etc.
- Valores escritos de maneira inadequada.

Nenhum destes pontos ocorreram no dataset. Mas com a ajuda de gráficos (boxplot, scatterplot e heatmap) além de outras estatísticas descritivas notamos que:

- Variável default e educação está desbalanceada.
- Existe possíveis outliers como uma renda de 446 milhões.
- Renda, divida_cc e outras_div estavam concentradas em um pequeno intervalo.

Tais inconsistências foram resolvidas da seguinte maneira:

- Educação foi mapeada para um novo intervalo com apenas três valores (antes era cinco).
- Aplicou-se o logarítmo em renda, divida_cc e outras div. Isso permitiu um maior espalhamento das variáveis e fez com que os outliers distantes ficassem agora bem próximos dos outros valores.

Amenizados tais problemas, identificamos que poderíamos criar uma nova variável (divida_cc + outras_div) que possuiria um poder descrito um pouco maior do que as duas em separado. Tal variável foi chamada de div.

Por fim, avaliando novamente os gráficos determinamos que as seguintes interações devem ser verificadas na etapa de modelagem por terem uma correlação alta ou por fazerem sentido. Tais interações são:

- renda:t_emprego
- outras_div:renda
- outras_div:divida_cc
- divida_cc:renda
- t_endereco:idade

Salientamos que educação deve ser avaliado com todas as outras, já que, não possuímos muitas informações sobre ela até o momento.

Modelagem

A modelagem levou em conta todas as suposições da análise exploratória e adicionou o seguinte fato: os dados foram padronizados (média 0 e desvio padrão igual a 1). Isso foi necessário para melhorar as acurácias dos modelos preditivos (verificamos que tal transformação causava melhoras).

Tendo os dados tratados, aplicamos métodos de seleção de variáveis, considerando a regressão logística, com distribuição binomial e função de ligação

logito. Essa é frequentemente usada para dados de credit scoring, semelhantes ao problema do trabalho. Abaixo há uma lista dos métodos que foram utilizados:

- Backwards via AIC (backAIC)
- Backwards via BIC (backBIC)
- Forward via AIC (forAIC)
- Stepwise via AIC (stepAIC)
- Lasso
- Ridge
- Knn

Após a criação destes modelos começamos a avaliar em separado os lineares generalizados (MLG) criados com AIC ou BIC e notamos os seguintes pontos:

- general e forAIC geraram os mesmos classificadores.
- backAIC e stepAIC geraram os mesmos classificadores.
- forAIC possui colinearidade, stepAIC também.

Desse modo, ‘cortamos’ alguns modelos e os reduzimos de sete para quatro:

- Backwards via BIC (backBIC)
- Lasso
- Ridge
- Knn

Sobrando apenas um MLG via BIC, aplicou-se uma abordagem híbrida ao problema, i.e, tendo agora um conjunto reduzido de variáveis significantes em backBIC, adicionamos aos poucos interações na finalidade de alcançar um modelo melhor. Essas adições foram avaliadas quanto a AIC e significância das covariáveis.

Além disso, é importante salientar que essa abordagem foi necessária pois quanto tentou-se adicionar diversas covariáveis e iterações juntas, ocorria o problema de separação perfeita (problema típico da regressão logística).

Abaixo há o resultado dos MLG’s mais interessantes das interações propostas da AED:

alternative1: default ~ t_endereco + t_emprego + div + renda + renda:t_emprego
alternative2: default ~ t_endereco + t_emprego + div + renda + renda:div
alternative3: default ~ t_endereco + t_emprego + div + idade + idade:t_endereco
alternative4: default ~ t_endereco + t_emprego + div + idade + idade:t_emprego

Com esses quatro novos modelos mais o backBIC nós realizamos uma análise de resíduos na finalidade de avaliar a sua qualidade. Pelos gráficos é possível concluir que todos os MLG's possuem uma nuvem aleatória em Resíduos x Valores Ajustados.

Contudo, ao analisar o qqplot pode-se verificar que apenas alternative1 e alternative2 possuem erros seguindo uma normal padrão. Os outros modelos acabam desviando nas pontas do gráfico.

Portanto, dos cinco MLG's sobraram apenas dois que somados com os outros classificadores geram estes modelos finais:

- Alternative1
- Alternative2
- Lasso
- Ridge
- Knn

Classificação

A classificação utilizou os cinco modelos da etapa anterior para avaliar qual seria um melhor classificador. Tal avaliação considerou a acurácia, sensibilidade, especificidade, AIC e AUC dos modelos. É importante salientar que dado o problema de desbalanceamento verificado na AED, realizou-se uma amostragem balanceada dos dados e os dividimos em dois conjuntos (treinamento/teste) com finalidade de avaliar melhor as medidas.

Ao final do cálculo das métricas notou-se que não era possível concluir com um alto grau de certeza qual classificador era melhor pois os valores eram muito similares.

Além disso, sabendo que tais medidas poderiam variar devido a variação da escolha do conjunto de treinamento e teste, extraiu-se uma média de AUC, sensibilidade e especificidade dentro de 50 valores calculados para cada modelo.

Desse modo, foi possível notar que knn, lasso e ridge possuíram um desempenho pior com relação aos modelos alternativos. Isso porque foram esses últimos que obtiveram um menor AUC, sensibilidade e especificidade no geral.

Portanto, nesta etapa podemos concluir que os seguintes modelos seriam adequados:

- Alternative1
- Alternative2

Salientamos que a escolha de qual utilizar para resolver o problema dependerá dos tomadores de decisão (tanto do cliente como da empresa consultora) pois

apesar dos modelos serem similares quanto a desempenho, eles consideram covariáveis diferentes. Logo, essas pessoas serão responsáveis por avaliar aquilo que faz mais sentido e é mais importante para o seu problema. Também salientamos que por não haver muitas observações (apenas 500) foi difícil os modelos capturarem bem as características de cada classe (principalmente de não default). Provavelmente se tivéssemos mais dados os modelos teriam métricas bem melhores.

Por fim note que para os métodos de regularização Lasso e Ridge, o programa R não dispõe de uma função igual ou similar a *summary()* para o modelo resultante desses métodos como ocorre para o modelo resultante de Backwards por exemplo. Dessa forma, a interpretação dos coeficiente e melhor conhecimento sobre o resultado do processo fica deficiente.

Interpretação de covariáveis

Considerando o Alternative1 e Alternative2, iremos interpretar as covariáveis usadas. Em resumo, podemos interpretar que o valor do coeficiente (β) indica que o acréscimo de uma unidade na respectiva variável explicativa resulta num acréscimo e^{β} à variável resposta.

Vale notar que em regressão logística, a interpretação dos parâmetros baseia-se em razão de chances, ou seja há esse efeito exponencial no acréscimo da chance de resposta y=1 para um acréscimo de 1 unidade na respectiva variável explicativa.

Dessa forma, em Alternative1, a variável t_emprego aumenta a chance de ser inadimplente em $e^{2.11}$ para o aumento unitário em 1 unidade, enquanto que renda e t_endereco têm efeitos menores, $e^{0.66}$ e $e^{0.48}$ respectivamente. Por outro lado, a variável div, que seria a soma das variáveis divida_cc e outras_div, tem um efeito de $e^{-1.62}$, o que diminui a chance de tornar-se inadimplente; e o termo de interação t_emprego:renda, efeito de $e^{-0.4}$. O intercepto do modelo é 3.6482, para uma variável binária, é um valor alto.

Em Alternative2, observamos que as variáveis t_endereco e t_emprego têm os menores efeitos com $e^{0.49}$ e $e^{1.39}$, respectivamente, ao passo que a variável renda aumenta a chance do indivíduo ser inadimplente em $e^{2.33}$ para um acréscimo em 1 unidade. Já para a variável div, soma das variáveis divida_cc e outras_div, o efeito é $e^{-0.82}$ indicando que diminui a chance do indivíduo ser inadimplente e o termo de interação div:renda tem um efeito de $e^{-0.55}$. Por fim, o intercepto do modelo é 1.4272.

Por fim levantamos algumas suposições (sobre ambos modelos) como:

- o perfil mais comum do cliente inadimplente pode ser baseado em uma pessoa mais velha, como acima dos 30 anos, a qual o padrão de vida tenha

aumentado, e com isso os gastos e despesas, como a aquisição da casa própria ou a geração de filhos, fatores que podem gerar dívidas.

- por outro lado, pode-se considerar que a base de dados é baseada em clientes que já tem um histórico no banco, e por isso, o banco empreste menos dinheiro àquele cliente que já tenha tido dívidas, tendo menos chance de se tornar inadimplente.
- sobre a interação t_emprego:renda, ocorre que se ambas crescem, isso contribui para a diminuição da chance de inadimplência.
- como visto no heatmap, dívida_cc e outras_dívidas têm boa correlação com renda, fazendo sentido a inclusão da interação, e o efeito de diminuição na variável resposta pode ser explicado pela questão do histórico do cliente como citado acima.

Conclusão

O problema de categorização em inadimplentes ou adimplentes, nos motivou a encontrar modelos através de métodos de seleção e regularização que nos indicassem e ajudassem a interpretar o papel de cada variável no modelo. Comparamos essas metodologias junto com o algoritmo de categorização Knn usando as medidas de acurácia, especificidade e sensibilidade. Com a médias de 50 resultados dessas medidas, concluímos que os métodos se equivalem, mantendo bons resultados com exceção de Knn, que obteve uma especificidade mais baixa.

Dessa forma, para poder entender melhor as covariáveis usadas nos modelos resultantes, pegamos o Alternative1 e Alternative2, ambos obtiveram resultados semelhantes, e chegou-se à conclusão de que o modelo a ser escolhido depende do quanto o banco prioriza as covariáveis, como t_emprego:renda ou div:renda.