Modelos de Collection Score

Renan Silva Ramalho Vilas Novas [1]
Aluno

Prof. Doutor José Mario Martinez [2]
Professor Orientador

Departamento de Matemática Aplicada, IMECC – UNICAMP, Campinas, SP

1. RESUMO

O objetivo deste trabalho é discorrer sobre meus aprendizados vinculados às atividades desenvolvidas na instituição *Itaú Unibanco Banco Múltiplo S.A.* ao longo do semestre.

Realizei um estágio na *Diretoria de Modelagem e Pesquisa* do banco, especificamente na *Superintendência de Métodos Quantitativos de Crédito*.

Minha principal função durante o semestre foi participar da elaboração de modelos de cobrança (*Collection Score*) para as mais diversas áreas do banco.

Estive em contato com todo o processo de construção de um modelo de cobrança, portanto buscarei aqui discorrer separadamente sobre cada uma das etapas da modelagem.

2.1. INTRODUÇÃO: O Modelo de Collection Score

Um modelo *de Collection Score* é um tipo de modelo construído buscando classificar o risco do cliente em termos de pagamentos futuros.

A população de entrada do mesmo é sempre um público já inadimplente. Logo, o objetivo do modelo é justamente fornecer uma medida (*score*), que ordene tais clientes inadimplentes numa escala de menor para maior probabilidade de recuperação/regularização da dívida.

[1] *vnovas.renan@gmail.com*
[2] *martinez@ime.unicamp.br*

Entre as principais utilidades de um modelo de *Collection Score* está a estruturação de uma régua de cobrança, que determine diferentes ações a serem aplicadas nos indivíduos inadimplentes, dependendo do número de dias de atraso do mesmo.

Uma régua de cobrança é uma estrutura fixa de políticas a serem aplicadas sobre o indivíduo inadimplente. Tais políticas são mais brandas ou mais agressivas dependendo do número de dias em atraso e do Score do cliente.

Figura 1: *Exemplo de régua de cobrança*

Outra grande utilidade do *Collection Score* seria na quantização dos descontos a serem aplicados pra indivíduos com maiores atrasos, buscando incentivar a regularização das dívidas do mesmo.

Públicos em faixas de atraso mais elevadas tendem a somente regularizar suas dívidas mediante renegociação da mesma. Portanto, utilizando o modelo é possível precificar o desconto máximo a ser concedido em cada uma destas renegociações.

2.2. MODELAGEM: Público do modelo

A primeira etapa da construção do modelo consiste na definição exata de qual é o público alvo do modelo, para o qual será gerado um *Score*.

O público escolhido deverá ser um subgrupo da população inadimplente cuja relação entre as variáveis explicativas e a variável resposta do modelo seja a mais fixa possível.

Buscando tal estabilidade no poder preditivo das variáveis do modelo, uma primeira quebra já pode ser realizada, a quebra de *buckets* de atraso.

Buckets de atraso correspondem a faixas de atraso na qual se encontram os indivíduos. Podemos definir por exemplo que existem 3 *buckets* que desejamos escorar, um de *5 a 30 dias de atraso*, outro de *31 a 60 dias de atraso* e outro de *61 a 90 dias de atraso*.

A motivação para tal divisão seria a de que em cada uma destas faixas de atraso, os motivos que levam um indivíduo a regularizar sua dívida são de natureza distinta. Entretanto dentro da mesma faixa de atraso, os indivíduos apresentam um comportamento bem homogêneo, o que permite o agrupamento.

Sendo assim, haverão modelos isolados para cada um dos *buckets*, que poderão depois ser realinhados e transpostos para uma mesma escala, buscando fazer com que mesmas faixas de scores possuam taxas semelhantes de regularização, independente do modelo.

Outro ponto importante de atenção na seleção do público a ser escorado é entender desde o início qual será a principal aplicação do modelo, dado que diferentes patamares de atraso serão abordados, dependendo desta aplicação.

Modelos com aplicações em réguas de cobrança normalmente buscam escorar faixas mais recentes de atraso, enquanto modelos de precificação de descontos normalmente escoram faixas de atraso mais elevadas.

Finalmente deve-se também determinar também quais serão as safras de público (meses de referência) utilizadas para a construção do modelo. O número de safras deverá ser suficiente para garantir que o modelo possua uma alta estabilidade temporal, e não descalibre tão facilmente após vários meses de uso. Entretanto muitas safras poderão também começar a interferir na capacidade preditiva do modelo nas safras atuais.

2.3. MODELAGEM: Variável resposta do modelo

Uma das principais técnicas de regressão utilizadas para este tipo de modelo é a *Regressão Logística*. Com ela, busca-se obter uma resposta binária para o problema a ser modelado.

Um exemplo de possível variável resposta para um modelo de cobrança é a *regularização da dívida nos próximos 30 dias*.

Caso o indivíduo regularize sua dívida no período de um mês ele é marcado como um indivíduo bom pelo modelo. Já indivíduos que regularizam após mais que um mês, ou não regularizam, devem ser marcados como maus.

A definição de como será marcada a performance de cada indivíduo inadimplente varia de acordo com a visão da empresa e de seu departamento de cobrança. A variável resposta do modelo deverá ser um reflexo do que a empresa enxerga como sendo um bom ou um mau indivíduo, do ponto de vista de cobrança.

Nem sempre tal simplificação de bons e maus é adotada, já que podem existir subgrupos na população inadimplente onde há uma maior dificuldade de categorização.

Pode ser que a empresa realmente considere quem regularizou a dívida nos próximos 30 dias um indivíduo bom. E pode ser que a empresa realmente considere maus aqueles indivíduos que não regularizaram ou demoraram *mais de 90 dias* para regularizar suas dívidas.

Entretanto para os indivíduos com regularização entre *31 e 90 dias*, o conceito de performance fica um pouco mais difícil de ser aplicado. Surge então a alternativa de declarar uma performance indeterminada para esta subpopulação do público, separando-a do restante do público para fins de modelagem.

O modelo seria então construído apenas analisando a correlação das variáveis explicativas com as populações boas e más. Entretanto como a população indeterminada também estará presente no público-alvo de *escoragem*, então a equação gerada deverá também ser capaz de ordenar bem esta subpopulação indeterminada.

O principal ganho na utilização da performance indeterminada é a obtenção de um modelo que consiga segregar melhor os bons dos maus indivíduos. Para isso abdica-se de parte do potencial de ordenação dos indivíduos com performance menos clara do ponto de vista da empresa.

2.4. MODELAGEM: Seleção de variáveis

2.4.1. Processo de Seleção: Visão Geral

O foco desta etapa é a filtragem de variáveis com maior poder explicativo com relação à variável resposta (performance). Visando garantir a estabilidade e a performance do modelo nas safras atuais, tais variáveis filtradas deverão ter um poder explicativo estável ao longo do tempo.

Normalmente centenas de variáveis são submetidas a uma rotina de testes estatísticos buscando definir um pequeno subconjunto de variáveis finalistas do modelo.

O conceito destas variáveis finalistas deve ser então bem compreendido, buscando assim garantir que o modelo seja realmente coerente com a realidade.

Um dos pontos de atenção nesta etapa é a relativização de variáveis absolutas, como por exemplo, valores monetários. Uma variável de saldo em atraso, por exemplo, pode até explicar muito bem o comportamento do indivíduo, entretanto devido à inflação ou a outros fatores macroeconômicos que norteiam o crédito, esta variável é pouco robusta do ponto de vista temporal.

Se relativizarmos por exemplo o saldo em atraso, criando uma variável que corresponda ao percentual do limite do cartão de crédito do indivíduo em atraso. A variável passa então a ficar mais robusta, normalmente até aumentando seu poder explicativo, já que ela perde parte de sua dependência das condições macroeconômicas do período.

Outro importante ponto de atenção é o cuidado de evitar variáveis que estejam correlacionadas com a própria utilização do modelo.

Um possível exemplo é o número de ligações que a central de cobrança do banco realiza para o indivíduo. Se a variável relaciona que quanto mais ligações o banco realiza, pior é o indivíduo, então quanto mais o banco liga para o indivíduo, pior fica seu Score.

Sendo que o Score do indivíduo piora, o banco tende a ser ainda mais agressivo em sua política de cobrança, ligando ainda mais para um indivíduo. Com isso, o simples fato de o banco realizar uma ligação para o indivíduo, pode transferir o mesmo indivíduo para uma diferente régua de cobrança, mais agressiva, que pode resultar em atritos desnecessários entre o banco e o cliente.

O último ponto de atenção seria um check-up final, questionando se existe alguma outra variável que a equipe de cobrança acredite ser discriminante, entretanto foi filtrada pelos testes estatísticos.

Com isso pode ser realizada uma repescagem de variáveis, garantindo que que o modelo realmente reflita a posição da equipe de cobrança frente ao público inadimplente.

2.4.2. Processo de Seleção: Métodos Estatísticos

Análise Univariada

Inicialmente buscamos verificar em algumas centenas ou milhares de variáveis aquelas que podem ser utilizadas para o público a ser modelado. Para isto verificamos basicamente o preenchimento destas, independentemente de seus significados.

Variáveis com mesmo preenchimento em mais de *98% dos indivíduos*, por exemplo, poderiam já ser excluídas das etapas de modelagem.

Análises Multivariadas

Através de uma análise multivariada deseja-se selecionar as variáveis mais correlacionadas com a performance, dentre as que menos estejam correlacionadas entre si.

Para isto as variáveis são categorizadas e observa-se a taxa de regularização dentro de cada uma das categorias criadas.

Para isto usa-se basicamente o conceito de *Information Value (IV)*, definido abaixo:

$$IV = (\%_{bons} - \%_{maus}) * \ln\left(\frac{\%_{bons}}{\%_{maus}}\right)$$

Sendo:

$$\%_{bons} = \left(\frac{Número\ de\ bons\ na\ categoria}{Total\ de\ bons\ no\ público}\right)$$

$$\%_{maus} = \left(\frac{Número\ de\ maus\ na\ categoria}{Total\ de\ maus\ no\ público}\right)$$

Um exemplo de utilização do *IV* é na categorização de uma variável de idade.

Inicialmente ordenamos o público pela variável idade e o dividimos em *50 grupos de mesma volumetria*. Observamos então qual a taxa de regularização dentro de cada grupo.

Quanto maior o *IV* de uma categoria, mais ela se difere da média do grupo. Categorias com mesma regularização que a taxa média de regularização do público recebem 0 de *IV*. Já variáveis com maior ou pior regularização possuem um *IV* positivo, diretamente proporcional ao poder da variável de distanciar o grupo da média do público.

O *IV* da variável é definido como sendo o somatório dos *IVs* de todas as categorias.

Após calcular o *IV* de todas as variáveis, fazemos um teste de correlação entre as variáveis e retiramos as de menor *IV*, dentro de um grupo de variáveis correlacionadas.

Definimos um grupo de variáveis correlacionadas, por exemplo, como um grupo de variáveis que possuam entre elas um *coeficiente de correlação de postos de Spearman* maior que *70% em módulo*, por exemplo.

O *coeficiente de correlação de postos de Spearman* é definido como sendo o *coeficiente de correlação de Pearson* entre as variáveis já categorizadas.

Sendo o coeficiente de correlação de Pearson:

$$\rho = \frac{cov(X,Y)}{\sqrt{var(X) * var(Y)}}$$

Sendo *X* e *Y* as duas variáveis das quais deseja-se verificar a correlação.

Selecionando apenas as variáveis de maior *IV* não correlacionadas agora fazemos uma análise mais fina da estabilidade do *IV* ao longo das safras e buscamos variáveis que realmente façam sentido para o modelo.

Caso ainda existam muitas variáveis após tais etapas de filtragem, pode-se definir um ponto de corte de IV, selecionando apenas variáveis, por exemplo, com um *IV superior a 5%*.

2.5. MODELAGEM: Regressão

Obtidas as variáveis finalistas é realizada então uma regressão linear utilizando um algoritmo *stepwise*, cuja escolha das variáveis preditivas a serem incorporadas ao modelo é realizada automaticamente acrescentando e retirando as variáveis finalistas uma a uma da equação final, a fim de se obter o melhor modelo possível com uma menor combinação possível de variáveis.

A etapa de regressão é um processo basicamente computacional, onde busca-se otimizar a combinação das variáveis utilizadas na equação final, obtendo um bom modelo com um número que não seja excessivo de variáveis.

3. VALIDAÇÃO DO MODELO

Após construída a equação final do modelo, testamos agora o modelo em safras mais recentes, fora do período de modelagem e verificamos quão bem ele consegue diferenciar o público bom do público mau.

Para isso utilizamos basicamente o teste *Kolmogorov-Smirnov (KS)* e o *coeficiente de Gini*.

O KS consiste basicamente na maior distância entre a curva de bons acumulados e a curva de maus acumulados, com relação ao score.

Sendo que a curva de bons e maus acumulados varia entre *0%* e *100%*, o máximo *KS* possível de ser obtido vale 1, que ocorre quando existe um escore n para o qual todos os indivíduos com score menor que n são maus e maiores que n são bons, ou vice-versa.

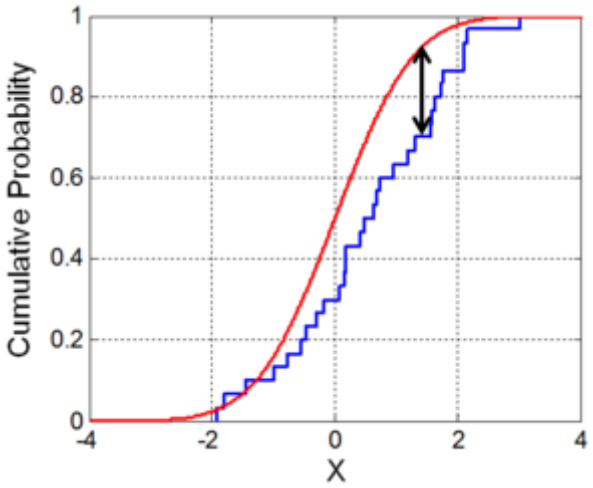

Figura 2: Exemplo do teste Kolmogorov – Smirnov entre duas funções acumuladas de probabilidade

Já o *coeficiente de Gini* busca verificar como está a ordenação nas faixas de score como um todo e não apenas na faixa onde há uma maior segregação entre a curva acumulada de bons e de maus. O *coeficiente de Gini* é o módulo da integral de uma das curvas acumuladas subtraída da integral da outra curva acumulada.

O *coeficiente* também varia entre 0 e 1, sendo que este vale 1 também exatamente quando o *KS* da mesma também é igual a 1.

Caso o modelo satisfaça todos as expectativas com relação à sua performance, ele agora já está pronto para ser utilizado.

Caso haja várias equações diferentes para um mesmo modelo, englobando cada uma um determinado subgrupo do público, é agora que estas devem ser realinhadas, para que fiquem em uma

mesma escala. Tal alinhamento permite que scores semelhantes em diferentes equações possam ser comparáveis.

Caso ainda haja algum descontentamento com relação à performance do modelo, deve-se voltar às etapas anteriores de seleção de público ou de seleção de variáveis, buscando encontrar algum ponto de melhoria no modelo, através de uma maior segregação do público de modelagem ou da construção de novas variáveis.

4. CONCLUSÃO

A criação de modelos de cobrança para escorar clientes é uma prática bastante problemática, dado a constante mudança do perfil do público inadimplente.

Um modelo muito bom de *Collection Score* pode buscar maximizar a regularização dos indivíduos inadimplentes, entretanto, se ele for bem sucedido, tal fato irá mudar o público inadimplente, tornando-o cada vez um público pior, o que torna sempre necessária a criação de um novo modelo de cobrança.

Além disso, os eventos de cobrança de uma empresa sofrem muito impacto de efeitos sazonais, após grandes feriados ou por exemplo o recebimento do 13º salário.

É um desafio bem interessante a construção de um modelo suficientemente robusto, que não descalibre à medida que a política de cobrança comece a ser implementada utilizando o score gerado.

Deve-se interpretar muito bem os resultados estatísticos obtidos e tentar justificar cada anormalidade sazonal. Uma greve dos Correios ou uma falha operacional podem causar inúmeros problemas se estiverem presentes no período da amostra de modelagem.

Todos os eventos anormais devem ser filtrados ou incluídos de uma maneira diferenciada no modelo, buscando maximizar o poder das variáveis explicativas de realmente retratar a realidade.

5. REFERÊNCIAS

http://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

http://en.wikipedia.org/wiki/Stepwise_regression

http://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

http://en.wikipedia.org/wiki/Lorenz_curve

http://www.plug-n-score.com/learning/characteristics-selection-using-information-value.htm