Nanodegree Engenheiro de Machine Learning

Proposta de projeto final

Gabriel Ramos Uaquim

04 de dezembro de 2018

Proposta

Histórico do assunto

Em 2014 com a recessão do país, a arrecadação caiu. Isso fez a Prefeitura Municipal de Salvador (PMS) voltar sua atenção para ações de cobrança. Dentre elas, uma reorganização das coordenações e setores dentro da Secretaria da Fazenda da PMS (SEFAZ-PMS), com isso foi criada a Coordenação de Cobrança, específica para pensar e executar a cobrança dos débitos dos contribuintes.

Alinhado a este raciocínio o presente trabalho busca soluções que visem otimizar o problema da cobrança de débitos. O uso de machine learning para tratar este problema não é novo ou recente. Diversas técnicas foram utilizadas [1] porém sempre foi dado preferência a técnicas lineares, devido ao seu alto poder de interpretação. Em [4] os autores tentam combinar estes métodos com técnicas com mais variância (árvores) para aumentar o poder de predição.

Em [3] e [4] se ressalva a importância da modelagem do problema, das variáveis escolhidas e do resultado a ser alcançado, pois aumentando essa especificidade ajuda a manter a estabilidade de modelos com maior variância como deep learning.

Porém a maioria desses técnicas e estudos são voltados para análise de risco e crédito no setor privado. Foi no início dos anos 2000 que uma técnica de cobrança no setor público ganhou notoriedade. Em Nova York, foi construída um sistema que utilizava um agente de aprendizagem por reforço para mapear a ação de cobrança do governo com o contribuinte e débito que deveria ser cobrado. O trabalho é descrito em [2].

O presente trabalho tenta adequar as técnicas usadas no setor privado para a análise da possibilidade de retorno de um débito do contribuinte.

Descrição do problema

Os débitos que o contribuinte pode ter possui diversas naturezas: oriundos de uma ação fiscal (ex: auto de infração), do não pagamento de um tributo devido, do não pagamento de um parcelamento e etc.

Dentre as fases do ciclo de crédito apresentados em [1], este trabalho tratará da *Collection Score*. Ou seja, daqueles débitos compreendidos no intervalo de atraso de 15 a 180 dias. Porém, como estamos tratando de governo, onde as pessoas tendem a atrasar mais as suas obrigações, e com a indicação da área de negócio, este prazo será aumentado para 30 e 360 dias.

Assim, problema é entender a relação entre as diversas características dos dados que são coletados como por exemplo (data de constituição do débito, valor, origem, região onde está localizado, se está em dívida ativa e etc) e uma chance de retorno desse débito. Além de entender essa relação, este projeto também tem como meta dar visibilidade a estes dados e propor soluções que otimizem o trabalho de cobrança.

Conjuntos de dados e entradas

A PMS mantém um data mart com os dados de produção da cobrança. Esses dados são carregados através de bases de diversos assuntos:

- Dívidas: são os débitos atrasados, acrescidos de juros, multa e correção monetária. Será a principal fonte de informação. Um débito só pode ser cobrado se já se transformou em dívida;
- Dívida ativa: algumas das dívidas vão para a dívida ativa para ser cobradas via execução fiscal. É um processo judicial moroso e com pouco retorno. Essa base irá ajudar a informar se a dívida ainda está sob a competência da coordenação de cobrança.
- ISS, IPTU, ITIV: são as informações da origem do débito. São informações como classificação da atividade econômica, notas fiscais emitidas, classificação do imóvel e etc. Não serão observadas as características específicas de cada imposto, apenas as suas naturezas.
- Arrecadação: informações de atualizações dos débitos, de recálculo das dívidas. Irá adicionar uma dimensão temporal extra à análise.

Essas informações provêm de diversos sistemas e fornecedores diferentes. A carga no datamart de cobrança facilita bastante o trabalho de obtenção desses dados.

Serão utilizados os dados referentes à 2016 e 2017.

Para modelar a variável alvo será feita a seguinte metodologia:

1. Para determinar se uma dívida é cobrável ou não será selecionado como cobrável todos os débitos pagos no intervalo de 30 a 360 dias.
2. Os débitos não pagos nesse intervalo serão selecionados como "Não Cobrável"
3. Para o modelo de Machine Learning:
    a. Cobrável == 1
    b. Não cobrável == 0

Descrição da solução

Como solução para o problema este trabalho irá ranquear os débitos. Do melhor cobrável para o pior. Um débito com "cobrabilidade" boa é um débito que foi pago em um período de 30 a 360 dias (Collection Score). Os débitos pagos em menos de 30 dias (Self Cure [1]) serão removidos da análise, pois são contribuintes que pagam suas dívidas voluntariamente.

Além de ranquear os débitos conforme a sua cobrabilidade, este estudo tentará dar visibilidade aos relacionamentos existentes entre as diversas características do débito e a variável dependente.

Dessa forma, ranqueando em porcentagens e demonstrando a importância de cada característica o presente trabalho irá ajudar a área de Cobrança a entender melhor seus débitos.

Modelo de referência (benchmark)

Os analistas e auditores do setor de cobrança utilizam um método ad-hoc para definir o que será cobrado. Uma lista é enviada à cobrança a cada dois meses. Dessa lista os débitos são priorizados conforme intuição. Infelizmente não é possível, até o momento, rastrear um pagamento a uma ação de cobrança. Assim, este trabalho irá buscar classificar a saúde do débito independentemente se houve cobrança ou não. O modelo de se determinar a saúde do débito, é chamado de *Collection Score,* e faz referência aos débitos pagos entre 15 e 180 dias (neste trabalho, por ser governo, 30 a 360 dias).

Acurácia Básica= Total Pagos / Total Débito

No conjunto inicial, com 2.481.582 de registros a acurácia básica é de 32%.

Usando o modelo básico GaussianNB, do sklearn, os valores de acurácia e fbeta (beta = 2) foram de 0.49 e 0.72 respectivamente.

Métricas de avaliação

A aferição da eficiência do modelo se dará a partir de quatro métricas:

1. Acurácia: métrica de fácil avaliação e principalmente fácil comunicação aos gestores da área de cobrança. Ela não será avaliada unicamente, pois as classes da variável dependente são desbalanceadas (78% e 32%), foi escolhida mais pela facilidade de comunica-la e explica-la.
2. Fbeta: para o presente trabalho temos recall e precision conforme descrito abaixo:
    a. Recall: um recall alto implica que poucos contribuintes que poderiam pagar foram deixados de fora. Em outras palavras, a Secretaria arrecada mais.
    b. Precision: uma boa precisão implica pouco desperdício no esforço da cobrança, aquilo que é cobrado é recebido.
    c. Beta = 2, pois o recall é uma métrica mais importante para a área de cobrança.
3. Curvas AUC: principalmente para comparar os diversos modelos, bem como também para comunicar graficamente o modelo escolhido.
4. Tempo de treino: o conjunto de dados é grande, e este trabalho deve ser repetível, a ponto de se construir uma funcionalidade para o próprio gestor treinar se for preciso. Tempos muito longos implicam em baixa usabilidade.

Design do projeto

Fase 1: Obtenção dos dados:

1. Reunião com a área de negócio para levantar parâmetros relevantes para a determinação das porcentagens e possíveis dificuldades a obtenção dos resultados;
2. Obtenção dos dados com o uso do Datamart da SEFAZ;
3. Os dados serão juntados e as chaves identificadoras dos contribuintes apagadas
4. Essa fase acontecerá no próprio datamart da prefeitura

Fase 2: Limpeza e arrumação dos dados:

1. Conversão dos campos para formatos adequados
2. Eliminação dos campos desnecessários ou fora do escopo

3. Uso de histogramas e boxplot para entender as variáveis quantitativas
4. Uso de Heatmaps para entender as variáveis qualitativas
5. Uso de árvore de decisão para entender a importância de cada uma das variáveis
6. Seleção preliminar das variáveis independentes

Fase 3: Construção do Modelo de dados:

1. Seleção de características:
    a. Para comparar as formas de seleção ou até a necessidade desta, será utilizado o modelo básico do sklearn GaussianNB. A forma que obtiver a melhor pontuação em acurácia será escolhida.
    b. Os modelos testados serão: PCA e SelectKBest do sklearn;
2. Seleção do modelo:
    a. O número de observações é grande. Cada ano tem em média mais de 1 milhão de observações, devido a isso o tempo de treino será também um fator relevante;
    b. Apesar da acurácia ser uma métrica muito mais fácil de comunicar à área de negócio, também usaremos Fbeta_score com beta = 2, pois o recall é mais importante para esse cenário.
        i. Recall: um recall baixo significa que a Secretaria deixaria de cobrar alguém que poderia pagar;
        ii. Precisão: uma precisão baixa implica a secretaria gastar recursos para cobrar alguém que não pagaria;
    c. Modelos testados:
        i. SGDClassifier
        ii. Logistic Regression
        iii. Kernel Approximation
        iv. Random Forest
        v. GradientBoosting
        vi. MLP Classifier
    d. Como são muitas observações métodos com muita variância terão um maior tempo de treinamento bem como um maior risco de variância.
3. Tunning:
    a. Será utilizado o GridSerchCV e o Pipeline caso necessário (ambos do SKlearn) para chegar ao melhor modelo dentro do modelo escolhido;
4. Determinação das probabilidades:
    a. Após obtido o modelo final, ele irá predizer as probabilidades apenas daqueles débitos que ainda não foram pagos;
    b. O conjunto com as probabilidade será salvo para ser usado no relatório final

Fase 4: Relatório Final

1. Construção do relatório com gráficos e insights analisando as porcentagens obtidas;
2. O relatório deve ser fácil de entender, pois será comunicado à área de negócio para tentar “vender” a solução.

Referências Bibliográficas

[1]. Forti, Melissa. **Técnicas de Machine Learning aplicadas na recuperação de crédito do mercado brasileiro**. 2018. Link: http://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/24653/Melissa_Forti_dissertacao.pdf?sequence=1&isAllowed=y

[2]. Abe, Naoki; Thomas, Vince; Kowalczyk et al. **Optimizing Debt Collections Using Constrained Reinforcement Learning**. 2010. http://www.cs.wayne.edu/~reddy/Papers/KDD10.pdf

[3]. Addo, Peter; Guegan, Dominique; Hassani, Bertrand. **Credit Risk Analysis Using Machine and Deep Learning Models**. https://www.mdpi.com/2227-9091/6/2/38/pdf

[4]. Galindo, Jorge & Tamayo, Pablo. (2000). Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. Computational Economics. 15. 107-43. 10.1023/A:1008699112516. https://www.researchgate.net/publication/5144412_Credit_Risk_Assessment_Using_Statistical_and_Machine_Learning_Basic_Methodology_and_Risk_Modeling_Applications

[5]. Dumitrescu, Elena et al. Machine Learning for Credit Scoring: Improving Logistic Regression with Non Linear Decision Tree Effects. 2017. https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=IAAE2018&paper_id=185