Avaliação Diagnóstica

Daniel Gardin Gratti
Instituto de Computação, UNICAMP
RA: 214729

I. Introdução

Este trabalho tem como objetivo uma análise exploratória inicial de uma base de dados de crédito sob a ótica de ética em aprendizado de máquina. Uma análise ética sobre os dados a serem trabalhados é um primeiro passo de alta importância para um bom continuamento da pesquisa, encontrar e mitigar problemas de vieses, falta de privacidade e injustiças logo nesta etapa inicial garante menos problemas em etapas posteriores, sem o risco de invalidar um modelo inteiro.

Um famoso caso de injustiça presente a uma das mais clássicas bases de dados, popularmente chamada de Boston housing prices [4]. Uma análise dos dados [3] feita por M. Carlisle mostrou que os dados continham uma informação racial de proporção de moradores negros em dada região. Esta variável era utilizada no artigo original para codificar guetos, regiões onde vivem pessoas pertencentes a um dado grupo minoritário e também considerar fatores de preconceito e segregação. Apesar dos dados em si não serem racistas, por refletirem uma realidade, o uso de informações que codificam a segregação para modelos de predição podem treinar um modelo que aprende o racismo estrutural, extrapolando o comportamento segregativo para outros contextos. Ao final, obtemos um modelo potencialmente racista que, apesar de outros fatores, penaliza regiões de alta concentração de negros apenas por si só.

Casos de problemas éticos não são incomuns e muitas vezes podem estar não tão óbvios, como apresentado no exemplo anterior. Neste sentido, uma análise estatística da origem, significado e relevância dos dados, assim como seu alinhamento com políticas de proteção de dados, como a LGPD, têm de ser alvos de preocupação de desenvolvedores de soluções com aprendizado de máquina. Este trabalho é uma análise inicial diagnóstica para a disciplina de tópicos Ética em Aprendizado de Máquina (MO810) lecionada pelo professor Dr. Marcos Medeiros Raimundo, no instituto de computação (IC) da UNICAMP.

II. Análise de Dados

Para este trabalho introdutório utilizaremos uma base de dados de crédito modificada da clássica German Credit Data [5], com o número de variáveis reduzidas e uma informação adicional: O nome do indíviduo. As variáveis da base de dados estão descritas na Tabela I.

Tabela I
Dicionário de variáveis do dataset modificado

Nome da Coluna	Tipo de dado	Descrição
Name	identificador	Nome do indivíduo
Age	numérico	Idade
Sex	categórico	Sexo
Job	numérico	Grau de habilidade
Housing	categórico	Tipo de propriedade de moradia
Saving account	categórico	Montante guardado em poupança
Checking account	categórico	Montante atual em conta corrente
Credit amount	numérico	Montante de crédito requisitado
Duration	numérico	Duração, em meses, do empréstimo
Purpose	categórico	Propósito do empréstimo
Risk	categórico	Indicador de bom ou mau pagador

Logo de início é possível visualizar que os dados contém uma informação que viola direitos a privacidade. A coluna Name é utilizada como um identificador da amostra, que permite sua identificação e rastreio de forma individual, essas informações, de caráter pessoal, como nome, sobrenome, identificação civil (RG ou CPF) e dados de geolocalização violam a privacidade do indíviduo. Estas informações, conforme a Lei Geral de Proteção de Dados Pessoais (LGPD) [2] determina, devem ser utilizadas com o consentimento do titular, assim como o conhecimento do próposito e uso de suas informações.

Como na análise de crédito, o treinamento e a fase de inferência independem de características pessoais identificatórias, o uso e armazenamento destas informações constitui uma potencial invasão de privacidade e está coberto pelo Artigo nº7 da mesma lei. Idealmente, estes dados deveriam ser anonimizados através de idenficadores numéricos aleatórios ou completamente removidos. Neste trabalho, a coluna foi completamente descartada tendo em vista sua não utilização para a determinação de bom ou mau pagador.

As demais informações foram mantidas e consideradas como importantes para a classificação. A base de dados possui um total de 1000 amostras, dentre elas, a maioria são de bons pagadores, como é mostrado na Figura 1. Os dados são desbalanceados com a probabilidade de um indivíduo ser considerado inadimplente de 30%, desconsiderando quaisquer variáveis. A razão do desbalanceamento pode ser explicada pela origem dos maus pagadores, são indivíduos que receberam crédito, ou seja, foram considerados bons pagadores a princípio, e não foram. Desta forma, apenas observamos antigos falsos positivos e os consideramos como representativos da distribuição de maus pagadores.

Como a origem dos dados e as distribuições reais não são conhecidos, testar a amostragem dos dados e sua representatividade real é difícil, senão impossível. Neste sentido, utilizaremos os dados supondo que representem as distribuições de bons e maus pagadores. Sob esta hipótese podemos verificar

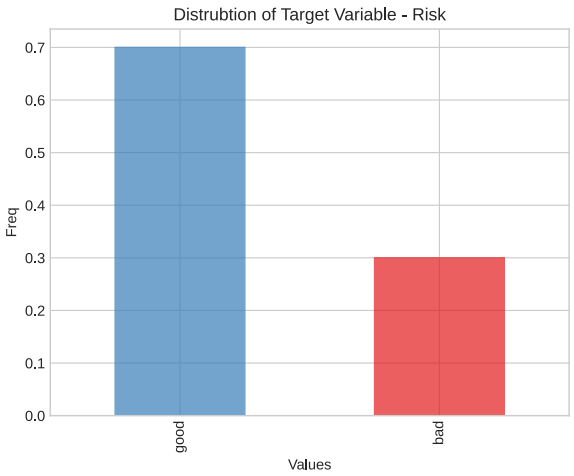

Figura 1. Distribuição de bons e maus pagadores.

as características individuais e as correlacionarmos com nosso alvo. A figura 2 mostra a distribuição de amostras por grau de habilidade da função. Apesar da maioria dos clientes estarem empregados, a proporção de maus pagadores em cada categoria de emprego é homogêneo, sendo 31.8% dentre os desempregados e 29.9% entre os empregados (28% dos trabalhadores de empregos de baixa, 29.5% dos trabalhadores de média e 34.5% dos trabalhadores de empregos de alta habilidade), indicando uma baixa correlação entre qualidade de emprego e qualidade de pagador.

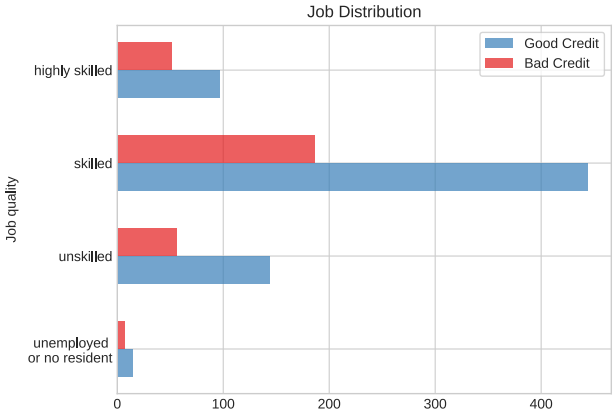

Figura 2. Distribuição de bons e maus pagadores por tipo de emprego.

Outra análise proposta é visualizar a distribuição de idade entre os pagadores, que não muda muito seu comportamento, conforme analisamos pela Figura 3 e possui distribuição com uma cauda direita longa (*e.g.* Lognormal, Beta, Weibull, Gamma). A fim de visualização de possíveis famílias de distribuições, realizamos o gráfico de Cullen e Frey, apresentado na Figura 8, em apêndice, para um guia inicial.

Testamos a hipótese nula de que nossos dados poderiam ter

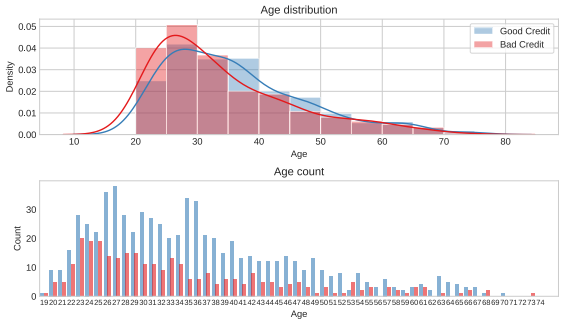

Figura 3. Distribuição de idade entre bons e maus pagadores

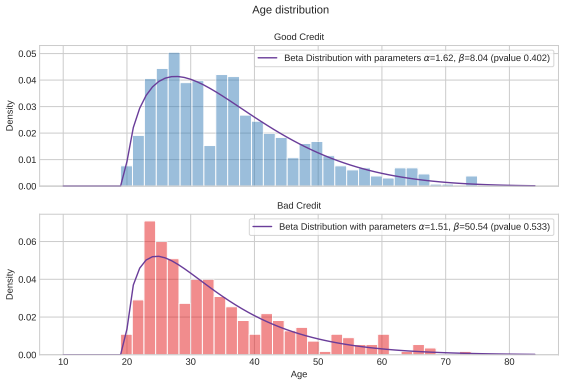

Figura 4. Distribuição de probabilidade Beta obtida a partir de regressão utilizando estratégia de bootstrapping. Ambas as distribuições estão deslocadas em 20 unidades em relação ao zero

sido amostrados por uma distribuição Beta, com parâmetros a serem encaixados por máxima verossimilhança. Para que o estimador se encaixasse bem aos dados realizamos uma estratégia de reamostragem por bootstrapping a partir dos dados disponíveis, obtendo parâmetros que melhor se encaixassem em diversos conjuntos reamostrados, cujo resultado é mostrado na Figura 4. A hipótese nula não pôde ser rejeitada devido aos p-valores elevados, o que indica que os dados podem ter sido amostrados a partir de uma Beta com tais parâmetros.

Apesar do bom encaixe das distribuições Beta encontradas para os dados, não é possível afirmar com certeza que esta é a distribuição que deu origem a ela. A determinação da distribuição de origem não pode ser feita com certeza e é possível que uma qualquer outra distribuição de probabilidade tenha gerado os dados, a análise garante apenas a possibilidade de modelar a distribuição de idade para o problema como uma distribuição Beta com os parâmetros encontrados.

III. Aprendizado de Máquina

Após a análise dos dados, feita na seção anterior, treinamos diversos modelos para avaliar o desempenho de cada um no problema de análise de crédito. Para isso, diversos modelos de classificação foram avaliados, dentre eles, Regressão Logística,

LDA, K vizinhos mais próximos, Árvore de Decisão, Naive Bayes, Floresta Aleatória, Máquinas de Vetores de suporte e XGBoost. Separamos o conjunto de dados em dados de treino, que constitui 75% das amostras originais e as demais 25% amostras foram separadas em um conjunto de teste.

Os dados então foram pré-processados utilizando normalização z-score e codificação one-hot das variáveis categóricas. Para avaliar a capacidade de generalização dos modelos, realizamos validação cruzada em 20 folds e as métricas de cada modelo estão na Figura 5.

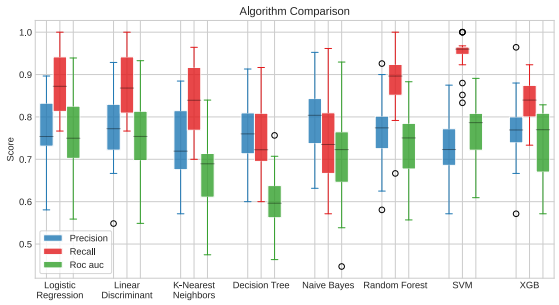

Figura 5. Performance de cada modelo por validação cruzada.

Após avaliação dos modelos, escolhemos a regressão logísitca para compor o modelo final, obtendo uma performance em conjunto de teste de 78% de precisão e 89.8& de recall. O desempenho do modelo não foi satisfatório, uma vez que a precisão, neste tipo de problema, é mais valorizada que o recall. O F-score, considerando importância de precisão 5 vezes mais que recall (proporção conforme a documentação dos dados indica [5]) é de apenas 78.4%. Uma vantagem do modelo escolhido, no entanto, é sua análise de explicabilidade. Como a regressão logística pondera sobre as features, é possível interpretar o valor de seus coeficientes como importância para a determinação de inadimplência.

A Figura 6 mostra os valores dos coeficientes obtidos após o treinamento do modelo logístico. Variáveis mais positivas influenciam positivamente na liberação de crédito, enquanto variáveis mais negativas diminuem a chance do empréstimo. As variáveis mais significativas, em módulo, foram de valor atual em conta, cuja influência é mais positiva para indivíduos com esta informação faltando e mais negativa para indivíduos com pouco dinheiro em conta. Logo em seguida, os propósitos que mais influenciam, atrás da quantidade em conta, são, positivamente, para rádio ou televisão, e mais negativamente para educação. Depois, a quantidade guardada em poupança impacta positivamente para ricos e negativamente para indivíduos com pouco dinheiro guardado.

Esta é uma forma de calcular a importância de cada variável, outros métodos estão disponíveis, como análises de ganho ao modificarmos variáveis individualmente, esta última, capaz de gerar explicabilidade em modelos mais complexos, e até modelos caixa-preta. A estratégia escolhida é uma forma mais simples de estimar a importância, pois seu valor está intrínsicamente ligado à probabilidade de saída e ao modelo.

Analisando ainda as importâncias obtidas, um debate ético que podemos propor é quanto a sua propagação de vieses. Apesar do modelo se encaixar nos dados de crédito, quaisquer vieses presentes nos dados, como a diferente ponderação em relação ao sexo, serão transferidos ao modelo. Isso impacta também pessoas cujo propósito de empréstimo é mais complexo, como o caso de educação, que é pesadamente penalizada, mas pode refletir um comportamento específico temporal, como uma crise.

Figura 6. Importância de cada variável, interpretada diretamente dos valores dos coeficientes aprendidos pelo modelo linear.

IV. Modelagem e Otimização

Nesta seção, focaremos em outro problema de classificação, dessa vez de um banco de dados do censo americano de 1994 e deseja-se determinar, através de dados pessoais sensíveis, como idade, sexo, escolaridade, se o indivíduo ganharia mais que 50 mil dólares anuais ou não [1]. Escolhemos um modelo linear de regressão logística utilizando o pacote de modelagem Pyomo[1] e solver Ipopt[2].

A modelagem original incluía variáveis categóricas não-ordenadas como atributos ordenados, como por exemplo, a coluna Country indica o país de origem do indivíduo, no entanto, sua coluna é numérica (ordem alfabética), o que causa injustiça ao atribuir valores maiores para certos países, e menores para outros. Consertamos este comportamento através de uma codificação one-hot e pré-processamento de outras variáveis, normalização de variáveis numéricas.

Esta mudança garantiu um aumento de performance em relação a modelagem originalmente proposta, obtendo um

[1]https://www.pyomo.org/
[2]https://github.com/coin-or/Ipopt

score ROC AUC de 0.91 e Recall total, em conjunto de teste, de 60% (10% de melhora em relação a modelagem inicial). No entanto, este resultado tem um grave problema ético, as métricas para o sexo masculino são significantemente superiores às métricas para o sexo feminino.

Em conjunto de teste, o modelo atingiu um recall de 61.9% para indivíduos do sexo masculino, enquanto obteve apenas 49.4% de recall para o sexo feminino. Esta diferença de mais de 10 pontos percentuais indica que nosso modelo aprendeu um viés de gênero, que podemos evitar, por exemplo, penalizando mais erros quando cometidos em indivíduos da classe menos favorecida. Retreinamos o modelo penalizando erros para o sexo feminino de forma que seja w vezes mais importante que o erro para o sexo masculino, conforme a função

$$\ell(\theta|X,y) = \ell(\theta|X_{\text{male}}, y_{\text{male}}) + \omega\,\ell(\theta|X_{\text{female}}, y_{\text{female}})$$

mostra. Os resultados estão mostrados na Figura 7, os pesos ω foram testados entre os valores 0.01 e 100, uniformemente distribuídos logaritmicamente neste intervalo.

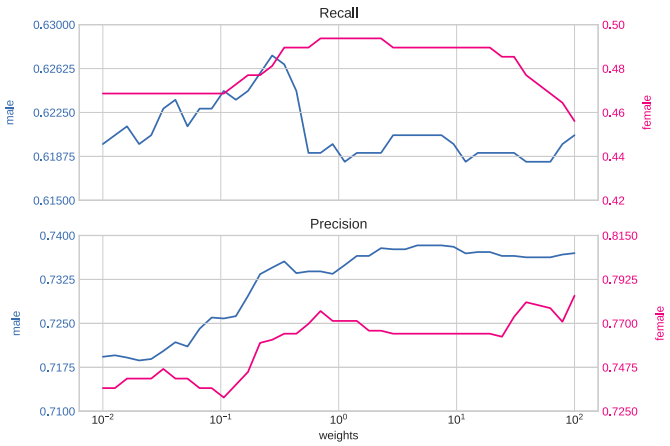

Figura 7. Recall e Precisão dos modelos com função de perda ponderada

Conforme pode ser visto, a estratégia não se mostrou muito efetiva em aumentar o recall nas amostras do sexo feminino, apenas em aumentar sua precisão. Este fenômenos talvez possa ser explicado devido ao grande desbalanceamento que a classe positiva possui. Como existem poucas amostras positivas, o modelo não consegue extrair variabilidade suficiente para gerar falsos positivos, mas gera falsos negativos com frequência, o que resulta numa queda de recall. É este o motivo pelo qual o recall é mais significante que a precisão.

Desta forma, somente esta abordagem não melhora o modelo, pois apesar de aumentar a precisão, o recall diminui conforme modificamos ω. Outras estratégias devem ser feitas para balancear melhor os dados enquanto deve garantir justiça.

V. Proposta de projeto

Nesta seção, discutimos brevemente um problema de ordem social para verificar vieses presentes em dados socioeconômicos, sociais e geográficos presentes na cidade de São Paulo. A partir dos dados disponibilizados, pela prefeitura da cidade de São Paulo[3], de informações locais socioeconômicas.

O objetivo do trabalho é, a partir somente de dados não-geográficos de diferentes locais da cidade, estimar o tempo médio de deslocamento e disponibilidade de transporte público do local de residência até o trabalho.

Referências

[1] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[2] Brasil. Lei nº 13.709, de 14 de agosto de 2018. *Diário Oficial [da] República Federativa do Brasil*, 2018.

[3] M Carlisle. Racist data destruction?, 2019.

[4] David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.

[5] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

[3] https://geosampa.prefeitura.sp.gov.br/

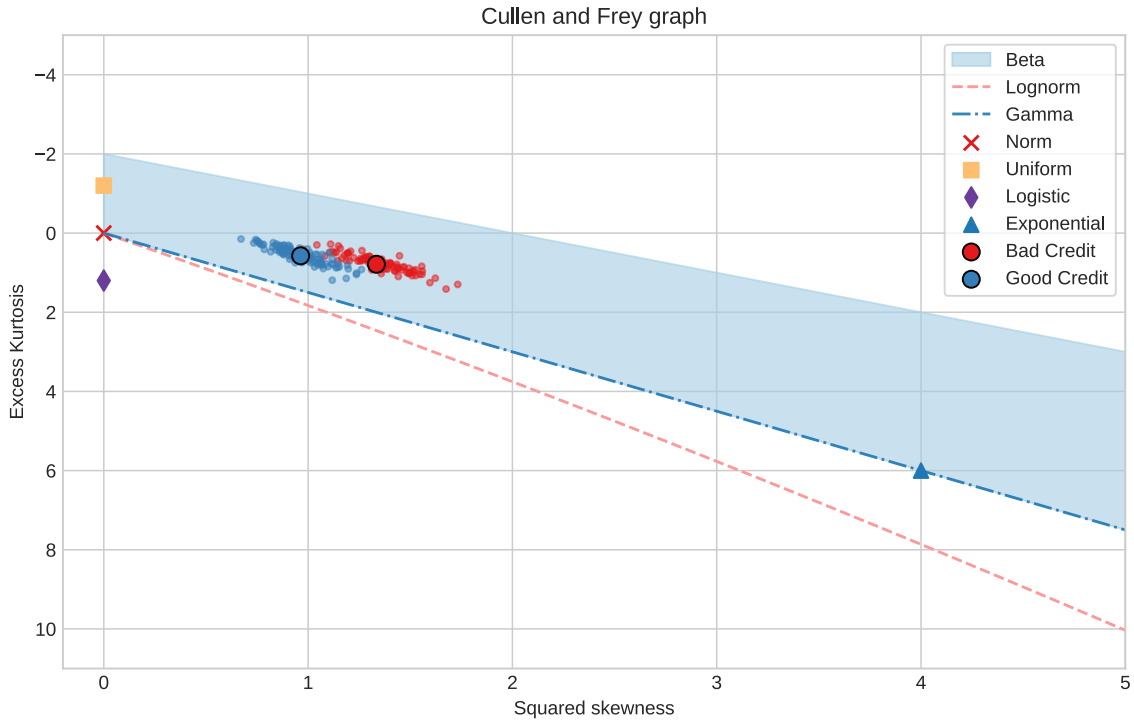

Figura 8. Diagrama de Pearson para diversas famílias de distribuições. As amostragens da base de dados estão marcadas na legenda, enquanto os pontos menores, de mesma cor, na vizinhança são obtidos através da reamostragem por bootstrapping. É possível ver que ambas possuem comportamentos parecidos e devem possui à família das distribuições Beta.