Ética em Aprendizado de Máquina

Fabio Fogliarini Brolesi
Instituto de Computação
Universidade Estadual de Campinas - UNICAMP
Campinas - Brasil
brolesi@gmail.com

I. Introdução

O presente trabalho tem como objetivo a apresentação de capacidade de entendimento e resolução de questões levantadas em uma avaliação diagnóstica do aluno.

II. Análise de dados

Os dados disponibilizados através do endereço https://drive.google.com/file/d/10JP23suA9vox-e37ya-tea16rJ0j-EDa/view?usp=sharing foram tratados no sentido de remoção de coluna de identificador e nome (`ID` e `Name` respectivamente) e de definição de variável categórica para colunas que de fato possuíam esta característica.

O conjunto possui 1000 registros, com as seguintes características:

- **`Age`** (numérico)
- **`Sex`** (texto: male, female)
- **`Job`** (categórico: 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)
- **`Housing`** (categórico: own, rent, or free)
- **`Saving accounts`** (categórico: little, moderate, quite rich, rich)
- **`Checking account`** (numérico)
- **`Credit amount`** (numérico)
- **`Duration`** (numérico)
- **`Purpose`** (categórico: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others)
- **`Risk`** (Variável alvo - Good ou Bad Risk)

A partir disso, avaliou-se cada um dos pontos abaixo.

A. Os dados apresentados tem problemas de privacidade? Quais?

Sim, os dados possuem problemas de privacidade. Eles não estão anonimizados e trazem uma coluna de identificador. Isso seria, do ponto de vista legal segundo a LGPD [1], um problema de privacidade: qualquer dado pessoal deve ser anonimizado para ser distribuído publicamente. No caso dos dados apresentados, é possível identificar a pessoa univocamente, qual seu patrimônio e qual a modalidade de residência que possui (se própria ou alugada).

Ainda assim, existe a possibilidade de os dados já serem anonimizados, com identificadores e nomes fictícios, o que não geraria um problema de privacidade, porém não está claro se este tipo de tratamento foi feito antes da disponibilização dos mesmos.

B. Sem considerar nenhuma variável independente, qual a probabilidade de alguém ficar inadimplente?

Dado o conjunto apresentado, temos que $P(Risk = bad) = 0,3$, ou seja, a probabilidade de alguém ser inadimplente é de 30%.

C. Qual a probabilidade de inadimplência dado que a pessoa tem emprego?

A respeito da probabilidade de inadimplência dado que a pessoa tem emprego:

$$P(i|e) = \frac{P(e|i) \times P(i)}{P(e)} \tag{1}$$

onde $c =$ inadimplência, $e =$ emprego

Considerando que inadimplência é onde a *feature* **Risk** é igual a *bad*, e que a pessoa ter emprego é onde a *feature* **Job** é igual a $1, 2,$ ou3, temos: $P(i|e) = 0,299591$, ou seja, a probabilidade de inadimplência dado que a pessoa tem emprego é de $29,96\%$ ou cerca de 3 em cada 10 pessoas.

D. Diga qual a distribuição das idades ao observar os dados? Como você fez isso?

A distribuição das idades ao observar os dados tem a característica de subir rápido e descer mais lentamente, lembrando uma distribuição Poisson ou uma distribuição Gama. Utilizou-se neste caso uma distribuição Gama, cuja equação, conforme , é dada por:

$$f(x) = \frac{(\frac{x-\mu}{\beta})^{\gamma-1} e^{(-\frac{x-\mu}{\beta})}}{\beta\Gamma(\gamma)} \qquad x \geq \mu; \gamma, \beta > 0$$

Onde γ é o parâmetro que representa o formato, μ é o parâmetro de localização, β é o parâmetro de escala, e γ é a função gama com a equação:

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$$

A curva foi identificada a partir do histograma de idade das pessoas participantes do dataset, com 50 bins, e utilizando a biblioteca `distfit` [2] do *python*. Dentre as várias distribuições possíveis em termos do padrão da biblioteca, a que foi mais aderente foi a gama.

E. Quais os parâmetros desta distribuição? Como você tem certeza disso?

Para identificar os parâmetros, apesar de a biblioteca `distfit` fornecer os parâmetros de melhor fit, utilizou-se o método `stats.gamma.fit` do módulo `stats` do pacote `scipy` [3]. O método utiliza o método `scipy.stats.rv_continuous.fit` que retorna estimativas de forma, localização e parâmetros de escala dos dados. O método de estimativa padrão para ele é Estimativa de Máxima Verossimilhança, mas o Método dos Momentos também está disponível para uso. Assim, os parâmetros para a função de distribuição gama são:

- $\alpha = 2.09770$
- $\beta = 8.00656$
- $\mu = 18.7506695811342$

Na Figura 1, à esquerda está o histograma das idades das pessoas do conjunto de dados e à direita, a modelagem gama considerando os parâmetros acima.

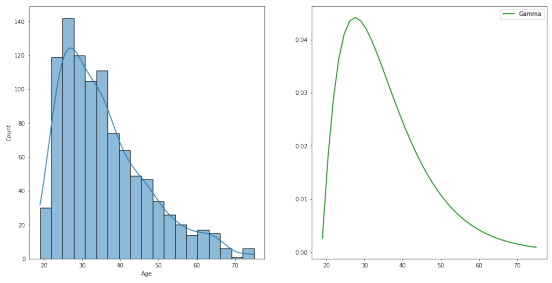

Figura 1. Distribuição de idades e distribuição gama

III. Aprendizado de Máquina

A. Qual o modelo de aprendizado de máquina que você acha mais adequado para os dados apresentados? Treine um modelo e reporte a performance

Para o treinamento, utilizou-se 3 modelos, baseados em levantamento feito por [4] e representado pela Figura 2: *Support Vector Machine*, *Logistic Regression* e *kNN*, com duas modalidades de *dummy features*: considerando todas as categorias e com $k-1$ dummies de k categorias, removendo o primeiro nível.

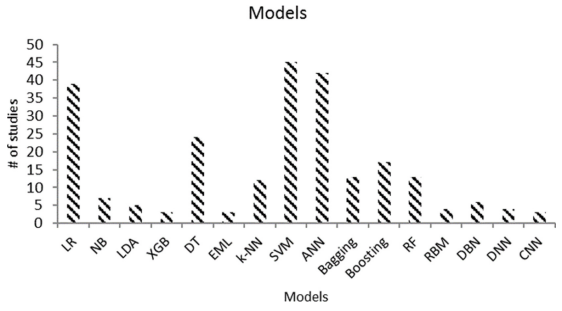

Figura 2. Um gráfico de distribuição de frequência dos modelos estatísticos e de aprendizado de máquina mais usados na pontuação de crédito conforme [4]

Tabela I
Matriz de confusão para o modelo de Logistic Regression

36	23
43	98

Tabela II
Principais metricas de classificação

	precision	recall	f1-score	support
0	0,46	0,61	0,52	59
1	0,81	0,70	0,75	141
accuracy			0,67	200
macro avg	0,63	0,65	0,63	200
weighted avg	0,71	0,67	0,68	200

Foi feita uma separação entre treino e teste, com 80% e 20% respectivamente, e o treino foi feito um downsample para que a diferença de 70% e 30% para a variável *target* **Risk** de *good* e *bad* ficasse 50% para ambos, mantendo a diferença para o teste. Também, para todos os casos, foi feito um tunning dos modelos, utilizando-se *grid search*.

O melhor modelo foi o de *Logistic Regression* com variáveis dummy completas. Para este caso, tivemos a matriz de confusão conforme a Tabela I e o relatório de méticas conforme a Tabela II.

B. Qual o mecanismo de controle de complexidade do modelo escolhido? Ele promove esparsidade no modelo?

Utilizou-se neste modelo, os dados categóricos em modo `dummy`, fazendo com que a esparsidade do dataframe fosse de $67,5\%$. Neste modelo, não houve um controle de complexidade, o que eventualmente pode resultar overfitting do modelo. Poderia ser aplicado PCA num pré processamento, antes da execução do modelo, porém, tal etapa não foi executada.

C. Quais são as cinco variáveis mais importantes para o modelo treinado? Como você estima as importâncias das variáveis, usando algum outro modelo ficaria mais fácil? Olhando as importâncias positivas e negativas encontradas, existe alguma variável que lhe levanta alguma preocupação ética?

Conforme avaliado, as cinco variáveis mais importantes estão descritas na Figura 3. Notou-se que características como ser homem, e ter uma casa própria além de ter uma conta poupança considerada rica são características que fazem com que a avaliação de crédito boa seja atribuída. Este caso é preocupante porque traz luz a fatos que geralmente excluem pessoas de estratos sociais mais pobres ou mulheres ou eventualmente outras minorias que podem ter possibilidade de receber crédito, mas pelo modelo podem ter este serviço negado.

IV. Otimização

O dataset apresentado consiste num conjunto com 32.561 linhas e 12 colunas da biblioteca *shap* [5] que contém dados de censo de população adulta. Estes dados originalmente provém

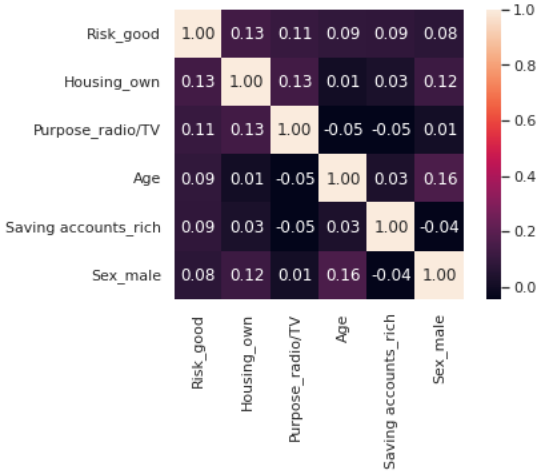

Figura 3. Correlação de Pearson considerando as 5 variáveis mais significativas para a variável resposta $Risk = good$

de [6] e conforme detalhado, é para predição se a renda excede US$ 50 mil/ano com base nos dados do censo.

A. Qual o problema de aprendizado de máquina modelado? Qual o modelo?

O problema é um problema de de predição baseado nos dados censitários. Foi utilizado uma solução utilizando a ferramenta *ipopt* conforme [7] para otimização não linear. Ocorre que os dados são relativamente antigos (mais de 20 anos) e podem conter imprecisões para os padrões atuais (desde questões que podem trazer problemas de privacidade, além de, ao final, o resultado (podendo ser positivo ou negativo) possivelmente não refletir a realidade.

B. Sabendo que l_h e l_m são os erros de aprendizado das amostras para homens e mulheres. Como você modelaria um problema para reduzir a diferença entre os erros de aprendizado?

Possivelmente, dado o cenário de avaliação se determinada pessoa recebe mais de US$ 50 mil, poderíamos utilizar outros atributos que não o de gênero, exatamente para não tornar o modelo enviesado por esta característica. Ocorre que o modelo pode não ter atributos suficientemente bons para uma boa modelagem. O que poderia ser feito é utilizar outros conjuntos de características que fossem consideradas boas para a análise. Importante lembrar que outras features também podem ter algum grau de correlação com o gênero da pessoa entrevistada, diminuindo o impacto desta característica, mas ao mesmo tempo, em algum grau considerando ela no modelo.

C. Essa é a melhor forma? Descreva de forma genérica alguma outra forma de fazer isso.

Não é a melhor forma. Avaliou-se que o viés (`BIAS`) é muito alto em relação aos atributos considerados para o modelo. Outros tipos de análise podem ser feitos ou outras features podem ser utilizadas. Podem existir outros tipos de modelos que não façam a diferenciação por gênero, mas por outras características que podem inclusive ser mais robustas dependendo do caso. É também necessário avaliar se o conjunto de dados tem features suficientemente boas para que um bom modelo (menos enviesado) seja desenhado.

Tabela III
MATRIZ DE CONFUSÃO

		Real	
		Positivo	**Negativo**
Predito	**Positivo**	VP	FP
	Negativo	FN	VN

D. A métrica de avaliação utilizada por grupo foi a recuperação (recall), qual a razão disso?

Dada uma matriz de confusão conforme III, temos a métrica de Recall é caracterizada por:

$$Recall = \frac{Verdadeiro\ Positivo}{Verdadeiro\ Positivo + Falso\ Negativo}$$

Também chamada de sensibilidade, probabilidade de detecção ou taxa de verdadeiro positivo, é a proporção de previsões positivas corretas para o total de exemplos positivos. Ou seja, ele penaliza (diminui o valor) quanto maior forem os falsos negativos, sendo ignorados os falsos positivos.

Avaliando que o recall para mulheres é menor, isso significa que a quantidade de erros para este estrato foi maior que dos homens.

V. LIMITAÇÕES

As limitações das análises foram as seguintes:

- Não foi feita uma análise mais profunda de que features do dataset de risco eram mais relevantes antes da execução do método de geração de *dummie variables*.
- O entendimento de otimização ainda não é suficiente para responder adequadamente todas as perguntas.
- Um estudo de bibliografia mais completa ajudaria a justificar a escolha da função gama para modelar a curva de idade do dataset.
- A ausência de domínio sobre LGPD e GDPR fez com que as justificativas de privacidade fossem escassas, e não houvesse menção a governança de dados.
- O desconhecimento dos datasets não possibilitou uma análise mais aprofundada com relação à variabilidade / outliers / missing values.
- Não foram testados modelos aditivos nem a avaliação da heterocedasticidade dos resíduos, mas apenas executados métodos padrão da biblioteca `distfit`.
- O downsizing não levou nenhum critério senão quantidade.

VI. PROPOSTA

Baseado em características de localização e dados censitários e/ou de saúde de um conjunto de pessoas, identificar se algum serviço básico é necessário em determinada região (postos de saúde, biblioteca, etc).

REFERÊNCIAS

[1] P. d. República, "Lei nº 13.709, de 14 de agosto de 2018," Aug 2018.

[2] E. Taskesen, "distfit - Probability density fitting," 1 2020.

[3] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[4] X. Dastile, T. Celik, and M. Potsane, "Statistical and machine learning models in credit scoring: A systematic literature survey," *Applied Soft Computing*, vol. 91, p. 106263, 2020.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 4765–4774, Curran Associates, Inc., 2017.

[6] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[7] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical programming*, vol. 106, no. 1, pp. 25–57, 2006.