Censo Estadunidense: Abordagens de Classificação

Lucas Camarino
lcealmeida@sga.pucminas.br
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil

Maria Luiza Lenti
maria.lenti@sga.pucminas.br
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil

Nathália Mascarenhas
nathalia.mascarenhas@sga.pucminas.br
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil

Raick Miranda
raick.miranda@sga.pucminas.br
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil

Túlio Brant
tbsguerra@sga.pucminas.br
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte, Minas Gerais, Brasil

ACM Reference Format:
Lucas Camarino, Maria Luiza Lenti, Nathália Mascarenhas, Raick Miranda, and Túlio Brant. 2023. Censo Estadunidense: Abordagens de Classificação. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUÇÃO

A análise da distribuição de renda em uma sociedade desempenha um papel crucial na compreensão das disparidades econômicas e na formulação de políticas públicas voltadas para a igualdade financeira. A base de dados *Adult Census Income*[1] é uma fonte rica de informações demográficas e socioeconômicas de indivíduos dos Estados Unidos, que permite a investigação de fatores determinantes na renda dessas pessoas. Este estudo tem como objetivo aplicar um algoritmo de classificação a fim de classificar as pessoas em duas categorias: aquelas que ganham anualmente menos ou igual a US$50,000 e aquelas que ganham mais de US$50,000.

O interesse primordial desta análise são indivíduos classificados na categoria de renda <=$50k. O foco na categoria de renda mais baixa é motivado pela necessidade de identificar as principais variáveis demográficas e socioeconômicas que contribuem para essa situação.

O restante desta análise está organizada da seguinte forma: na seção 2, será descrita detalhadamente a base escolhida, apresentando as instâncias, atributos e quais os valores que cada atributo assume. Na seção 3, serão apresentadas todas as etapas de pré-processamento utilizadas e os métodos escolhidos. Por fim, na seção 4, serão apresentados os resultados da análise e a discussão das principais descobertas. não

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 DESCRIÇÃO DA BASE DE DADOS

O conjunto de dados escolhido para análise é o *Adult Census Income*, que é composto por 14 atributos e um total de 32.561 instâncias. O principal objetivo deste conjunto de dados é realizar a previsão de se a renda de um indivíduo excede US$50 mil por ano com base em informações demográficas e socioeconômicas coletadas no censo.

Cada registo contém as seguintes informações sobre um indivíduo:

- ***Age***: Idade do Indivíduo.
  - Numérico.
- ***Workclass***: Classe de trabalho do indivíduo.
  - *Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.*
- ***Final Weight***: O número de pessoas que o censo acredita que o registo representa.
  - Numérico.
- ***Education***: Grau de instrução do indivíduo.
  - Categórico ordinal.
- ***Education-num***: Mais alto nível de escolaridade alcançado.
  - Numérico.
- ***Marital status***: Estado civil do indivíduo.
  - Categórico nominal.
- ***Occupation***: Profissão do indivíduo.
  - Categórico nominal.
- ***Relationship***: A relação familiar em relação aos outros.
  - *Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.*
- ***Race***: Descrições da raça de um indivíduo.
  - Categórico nominal.
- ***Sex***: Sexo biológico do indivíduo.
  - Categórico.
- ***Capital-gain***: Ganho de capital para um indivíduo.
  - Numérico.
- **Capital-loss**: Perda de capital para um indivíduo.

  - Numérico.
- ***Hours-per-week***: Horas de trabalho do indivíduo por semana.
  - Contínuo.
- ***Native country***: País de origem do indivíduo.
  - Categórico nominal.
- ***Income* (rótulo)**: Se o indivíduo ganha ou não mais de 50k anualmente.
  - <=50k ou >50k.

3 ETAPAS DE PRÉ-PROCESSAMENTO

Foram utilizadas duas etapas de pré-processamento, conversão numérica e balanceamento dos dados. A Figura 1 representa a ordem das etapas de pré-processamento utilizadas. Primeiramente os dados foram convertidos para numérico, depois foi feita a divisão entre teste e treino e, por fim, o balanceamento dos dados.

Conversão Numérica

A base de dados utilizada possui atributos categórico/nominais, dessa forma é necessário converter esses dados para valores numéricos. Os dados convertidos foram:

- Work class
- Education
- Marital status
- Occupation
- Relationship
- Race
- Sex
- Native country

Primeiro, foram identificadas colunas categóricas no conjunto de dados que precisavam ser tratadas. Em seguida, para cada uma dessas colunas, foi utilizada a técnica *Label Encoding* (Codificação de Rótulo). Isso significa que um número único foi atribuído a cada categoria dentro dessas colunas, de forma que os dados agora pudessem ser representados numericamente.

Após a codificação de rótulos, foi aplicada a técnica *One-Hot Encoding* usando o *ColumnTransformer*. Essa técnica transforma as colunas categóricas em um conjunto de colunas binárias (0 ou 1) para representar a presença ou ausência de cada categoria. Isso é feito para evitar que o modelo de aprendizado de máquina interprete erroneamente relações ordinais entre as categorias.

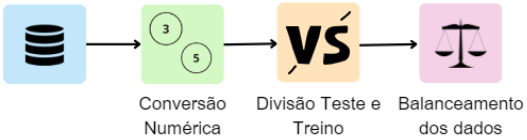

Figure 1: Ordem das Etapas de Pré-processamento

Divisão Teste e Treino

A divisão entre teste e treinamento foi feita antes do balanceamento. Foi utilizada a biblioteca `sklearn.model_selection`, que contém a função `train_test_split`. O tamanho escolhido para teste foi de 15%, ou seja, 85% das instâncias foram utilizadas para treino.

Balanceameneto dos Dados

O balanceamento de dados foi utilizado, já que no total, 24.720 dos indivíduos foram classificados como <=50k e somente 7.841 foram classificados como >50k. Foram utilizados 4 diferentes métodos de balanceamento, dois métodos *Over sampling* e dois *Under sampling*.

Primeiro foi utilizado o método *Random Over Sampling*, ele gera novas amostras da classe minoritária por meio de reposição de amostras. Também como método *Over*, foi utilizado o *Smote*, que ao invés de gerar novas amostras aleatoriamente, seleciona cada amostra da classe minoritária e introduz dados sintéticos que conectam a amostra minoritária com seus vizinhos mais próximos. Os vizinhos dos k vizinhos mais próximos são escolhidos aleatóriamente.

Os métodos *Under* foram o *Random Under Sampling*, que deleta amostras aleatórias da classe maioritária, e o *NearMiss*, que utiliza KNN para readuzir as amostras. Existem 3 versões de algoritmos *NearMiss*:

(1) Seleciona as amostras positivas para quais a distância média para as N amostras mais próximas da classe minoritária for menor.
(2) Seleciona as amostras positivas para quais a distância média para as N amostras mais distantes da classe minoritária for menor.
(3) É dividido em dois passos. Primeiro, para cada amostra negativa, seus M vizinhos mais próximos são mantidos, depois, as amostras positivas selecionadas são as quais a distância média para os N vizinhos mais próximos for a maior.

Para este estudo, os métodos *Over* são mais interessantes, visto que aumentarão a quantidade de indivíduos com >50k e manterá as mesmas instâncias para os classificados como <=50k, deixando-as mais completas. Contudo, serão discutidos os resultados de todos os métodos para a avaliação.

4 RESULTADOS E DISCUSSÕES

O algoritmo utilizado para classificação foi o *CART*, um dos algoritmos mais comuns para a construção de árvores de decisão em problemas de classificação e regressão. A Figura 2 apresenta as métricas de avaliação da árvore sem balanceamento e com os 4 tipos diferentes de balanceamento.

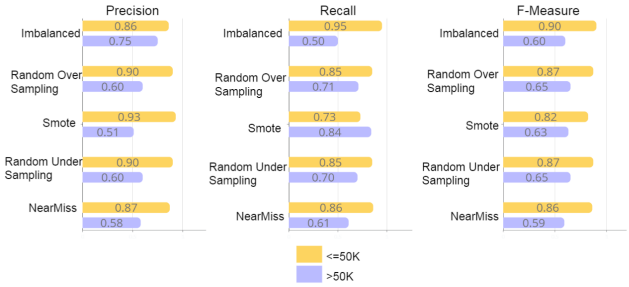

Figure 2: Métricas de Avaliação Com e Sem Balanceamento

Com relação aos valores de *Recall*, que mede a proporção de exemplos positivos que foram corretamente classificados pelo modelo, os melhores resultados, tanto para instâncias classificadas como <=50k, quanto para >50k, foram do *Random Under Sampling* e do *Random Over Sampling*, com *recall* maior que 0.7 nos dois casos. Essa métrica é essencial para esta análise, já que o foco é analisar todos os indivíduos que possuem renda <=50k, mesmo que algumas pessoas que recebam mais também sejam examinadas, ou seja, a taxa de falsos negativos deve ser baixa. Olhando apenas para os valores, a árvore sem balanceamento apresentou um melhor resultado, com *recall* igual a 0.95, no entanto, como a quantidade de amostras é muito distinta, essa árvore foi descartada, restando o resultado da *NearMiss* com 0.86.

Observa-se que, com relação a precisão, que mede a proporção de exemplos classificados como positivos pelo modelo que são realmente positivos, os melhores resultados foram o *Random Under Sampling* e o *Random Over Sampling*, que obtiveram boas precisões, tanto para <=50k, quanto para >50k, diferente dos outros, que ofereceram resultados muito bons para um, e ruins para o outro. Como o atributo mais interessante para o objetivo da análise é o <=50k, o melhor resultado foi o obtido pelo *Smote*, que apresentou maior precisão, porém, caso não fosse, os melhores seriam os do *Random Under Sampling* e *Random Over Sampling*, já que a árvore desbalanceada é desconsiderada.

Na métrica *F-Measure*, que é a média harmônica entre precisão e *recall*, os melhores resultados foram os do *Random Under Sampling* e *Random Over Sampling*, que apresentaram resultados iguais a 0.87.

Em suma, todas as árvores, balanceadas ou não, apresentaram melhor resultado para instâncias classificadas como <=50k, que é mais relevante, em comparação com as que foram classificadas como >50k. As melhores métricas para a opção mais relevante foram as do *Random Under Sampling* e *Random Over Sampling*, com foco no *recall*, que é muito importante para esta análise.

Por fim, para escolher o melhor método de balanceamento, foi levado em consideração o fato de estratégias de *Over Sampling* influenciarem apenas na classe minoritária e as de *Under Sampling* na classe maioritária, optando por descartar o *Random Under Sampling* e *NearMiss* e selecionar o *Random Over Sampling* para garantir mais informações da classe relevante.

5 CÓDIGOS DESENVOLVIDOS

https://github.com/raickmiranda/DataScienceAI_TP.git

REFERENCES

[1] 2023. Adult Census Income. https://www.kaggle.com/datasets/uciml/adult-census-income