Handling Class Imbalance with oversampling for predicting annual income: Duplication vs. synthetic data

Manar Attar (2631465), Majd Al Ali (2659280), Imane Akhyar (2667107), Duaa Ashtar (2696234) and Raihan Karim Ishmam (2694796)

Abstract

According to Pew Research Center, 71% of the world population has low or poor income(1). This huge imbalance in income distribution makes it hard for machine learning algorithms to make accurate predictions. In this paper, we decided to train different classifiers using an imbalanced dataset and optimize performance by attempting to balance the data. The aim is to predict the income, for which we use a dataset from the U.S. Census Bureau(2). To deal with the imbalance, two variants of oversampling: random oversampling (duplication) and synthetic oversampling (SMOTE) are compared to see which method performs better for different classifiers. the comparison is done by examining performance metrics such as f1-score and roc-AUC score. We observe an improvement for these scores using both oversampling methods, with slight advantage when using SMOTE-oversampling.

Keywords: Class Imbalance, Predicting Income, Deep Learning, Classification Algorithms, Machine Learning Models, Decision Trees, Random Forest, Support Vector Machine, Naive Bayes, K-Nearest Neighbours, Logistic Regression

1 Introduction

1.1 Motivation

One of the significant elements of our daily lives is our financial activity and status. From travel to food to shelter, all are highly dependent on and run by our finances (payments, rents, bills, etc.). Moreover, the most common source of providing this financial support is their income if we focus on adults. This makes income prediction an area of interest for many organizations and researchers. One of the most reliable datasets that we can use, including the data relevant to the

income, is the census dataset. To find a dataset of this type, we turned to two very well-known web repositories: Kaggle and the UCI - Machine Learning Repository. We initially aimed to find our datasets from the ongoing competitions in Kaggle, but we could not find an appropriate one matching our project. So we moved to the UCI repository, where we found a data set that contains weighted census data extracted from the 1994 and 1995 population surveys conducted by the U.S. Census Bureau(2). The data set contains demographic and employment-related variables, which would provide us with a set of reasonably relevant features.

1.2 Paper outline

The essential element of this project is using machine learning to predict the annual income of the various instances (adults). We decided to approach the problem as a classification problem rather than a regression one, as this would let us focus more on the statistical aspects of the performance. Choosing regression might have led to more detailed predicted values for the income, but that might introduce issues with the overfitting when we try to improve the accuracy, so we opted to go with classification. The machine learning classifier that we are developing is designed to do binary classification, again due to the reasons above. We chose to go with a specified value of (50,000 USD) to split the annual income range into two classes, so the instances are classified as, with an annual income either greater than 50,000 USD or less.

The dataset that we are using is comparatively significant, with around 32,500 instances. We are planning to use several different training models like K - Nearest Neighbours, Logistic Regression,

Handling Class Imbalance with oversampling for predicting annual income Duplication vs. synthetic data

Come refere (Smith) per de la companya del companya del la companya del companya de la companya de la companya del company

Construction (2010), Margin and Establishing Systems alloyer (2011) (1911)

Construction (2010) (1911) and finisher factor physics (2011) (1911)

Solvey Construction (2011)

Solvey Construction (201

serverability described in the qualitation of the product of the control of the point of a server and could be a first and the point of a server and could be a first and the could be a first and t

readons bearing desirble to compare their performances for making the observation of a regardered fordations or the observation of their subsense to their sections or consideration to believe the bearing date and observe the contract of the disease or the blackets performances on the best-like.

The person tolerant counts of 2 classes, solving the legal disabilities of 2000 intervalvies. The legal disabilities have been a serviced to the intervalvies does included the intervalvies observed and applications of the Intervalvies observed and a service of the Intervalvies observed and intervalvies of the Intervalvies of

International Conference (i.e., the high-inverse date.

Fills 7 file on around internaty (i.e., the high-inverse date.

Fills 7 file on around internaty - Secretic international conference (i.e., the high-inverse date date.) In the conference of the file of the conference of the fills of the fills of the conference of the fills o





oth is not in content, and it flowers, the explanate recipil, [6] as internatifica-ria longs a milks spine. The left spine is inconcepting, and the less stated in text are. The annual filliant tracket of memorphic, [6], and the properties of the inconcepting, and the second longer interprets the perfections of the in-liability self-bills, including the properties of the second participants. We have been also and participants. We have in the community of the second participants.

13. Manage union.

Technologicals interrequestion, our estimate pretraining of the control of

SZ Symptoms

Has Asserber's Regards to passed the convert, and the larger does inverted assemble from the cost of the changes which has also the cost of the changes, which was also the cost of the changes of the ch ment mention (AM), the said that these mentings, the control indexes consumers to said your mention (also recovered to the control indexes consumers to share control indexes that a part of the consumers that is shared to control and the consumers that is shared to the consumers that is shared to the consumers to the shared to the consumers to the control indexes of control indexes of the control indexes of control indexes of the control indexes of the

marky self a rise, relations in representations of the control of

In programs. Boardy is declarate the solutions of a factor of an interpretation flow some statement of an experiment of an experiment of the solution of the s

F Statement April 2015 Egyppisanit in Sprinkers 52 Statemengher (606)

The Brazant confine New York The Brazant Conf



C2 Summer bears: In advanced page of the control of

Impried Charallinghins. Der bei von versch in verschundig genet, der und dir der ten, verderig der verschundig genet, der und dir der ten, verbrieg der verschund der vers

A few former? Application for any electronic responses to complete the control of the control of

Paper de Trades de la Company de la Company de Company

"girl" and "colongs" orbits, and Albahad Burtle. Bur X from the assume temperature factories. Space below, where "group is the probability of a colony (21).

 $\{(a+1): \sum_{i=1}^n a^i > i\}$

between 1 and 20 years beyond the best increases are experiented by the contract that is all the contract to t The try supplies and the aftern further in the control of the company of the proble of the company of the proble of the company of the proble of the company of the control of the company of the control of the control

Since 1 is \$\int_{\infty}\eta^{\text{\$\mu}}(\text{\$\mu})^{\text{\$\mu}}}(\text{\$\mu})^{\text{\$\mu}}(\text{\$\mu})^{\text{\$\mu}}(\text{\$\mu})^{\text{\$\mu}}}(\text{\$\mu})^{\text{\$\mu}}(\text{\$\mu})^{\text{\$\mu}}}(\text{\$\mu})^{\text{\$\mu}}(\text{\$\mu})^{\text{\$\mu}}}(\text{\$\mu}

"Williams" (sag." Sagat) and the reducit value in consolidable bett assumed to see obtained using all a lower than agreement agree of the base. It is page, a detailed to the second of the lower and to the second of the lower and to the second of the lower and the lowe

the foundation and separated feating algorithms found colleges flowering department and colleges flowering the feature flowering the colleges for the flowering flowering the colleges flowering flo

bill Tigenting from host to large. The account of the probabilities in large to the same displace to the probabilities and the probabilities of the large to the same displace to the probabilities of the large to the same displaced to the probabilities of the large to the same displaced to the same of the large to th

District and the next seed seed seed

principal de la lors commission de la lors contrata de la lors commission de la lors com





12. Langing braids
to institute provising the symbols make it is a similar to be a similar to

LE Bearmenmenthiques in polement in the critical comparison, one permission of the critical comparison, one permission of the critical comparison of the critical control of t

NAME OF

Addition are referred research of continuous analysis, broad of a failurating for colors of a continuous analysis, and a failurating for colors of continuous analysis continuous analysis of the colors of the colo

Payer of revision and Applications
Representation
The goals allowed to the state of the stat

Translation
In resolution (in resolution) and its manufaction and its good for the resolution of the resolutio

ACC same for differentials except the contribu-fered and liquid registation, and both stages and fine search of the date of the conducting to different distincts that are below to the displacement. Evaluation, of the contribution of the ACOCOT is compared to consider the page 14th and below the desirable to consider the contribution of the COCOT of the ACOCOT of the COCOT of the performent. Let marking mar reportations, on solutional impercutation in performance with sales (ACCE) for self-timentals, southing from a distance or the market of table regulate that incorporate to associational project in the requirily They ded in temporal to the profile rate.

(far, ded in temporal to temporal to the first rate.)

(inclinate parties for all the model, say to the control of the control o

This has the country of the country

[4] I is to lead of the property of the proper

(i) 1.1 male of the P broad in terming and the read broader management and read in terminal management and proof to be a second or the proof t

The American Company of the Company

(b) it means por t. bland "may many one transcript former transcript com-ingle contract means, it was

THE PARTY NAMED AND POST OFFICE ADDRESS OF THE PERSONNEL PROPERTY.



TI

Openio di Conferenza non MATTICO











