Documentation

Vaccination Coverage for zero dose children

NAVEEN NANDAKUMAR

Professors:
prof. SUMEET AGARWAL
prof. MIRA JOHRI
prof. AADITESHWAR SETH

Contents

1 INTRODUCTION

Using the DHS NFHS dataset, the objective is to recode the dataset to a usable form(for modeling) and get started on developing a computational statistical model of childhood vaccination outcomes.

GITHUB: `https://github.com/masternaveen123/vaccination_coverage`

With the email, I have attached the following files: Script for recoding the dataset(Recoding.ipynb) Metadata of recoded variables(DPT vs Penta - Variables.xlsx) `Modeling approach(modelling_zero_dose.ipynb)`

Definitions:

- Zero doses are those data points without the first dose of DPT or pentavalent vaccines.
- Zero doses are those data points without the first dose of DPT or pentavalent vaccines.
- Fully vac are those with doses of BCG + OPV 3rd dose + (DPT 3rd dose— PENTA 3rd dose) + MR1 [a total of 8 doses]
- Under vacc is the inverse of full vac.
- NULL refers to empty or missing values

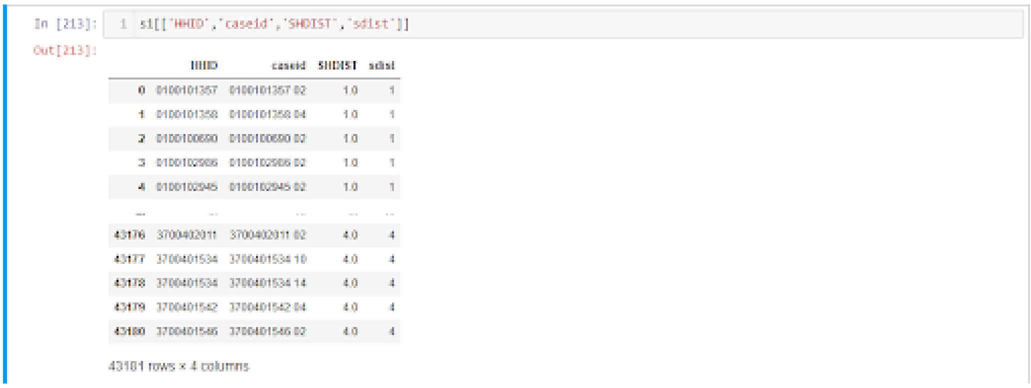
2 SCRIPT FOR RECODING THE DATASET

The script is written in Python, and it recodes the NFHS 5 dataset to be used for modeling.
The script takes NFHS 5 data along with a binary flag as input. If the flag is true, it converts the dataset to binary values only, otherwise, some values are converted to binary whereas some columns contain categorical values. The dataset is filtered to only have records of kids between the age range of 12-23 months(including).

Variables used as inputs:

- `Categorical_eleminate_flag (binary)`
- Children(location to DTA file containing children's records)
- Household(location to DTA file containing household records)
- Individual(location to DTA file containing individual records)

We have merged the household and children's record to get the household details for every child datapoint.
Verifying if there is any mismatch during the merging of household records and children's record



```
In [213]:  1  s1[['HHID','caseid','SHDIST','sdist']]
Out[213]:
                   HHID          caseid  SHDIST  sdist
          0  0100101357  0100101357 02     1.0      1
          1  0100101358  0100101358 04     1.0      1
          2  0100100690  0100100690 02     1.0      1
          3  0100102986  0100102986 02     1.0      1
          4  0100102945  0100102945 02     1.0      1
        ...         ...             ...     ...    ...
      43176  3700402011  3700402011 02     4.0      4
      43177  3700401534  3700401534 10     4.0      4
      43178  3700401534  3700401534 14     4.0      4
      43179  3700401542  3700401542 04     4.0      4
      43180  3700401546  3700401546 02     4.0      4

     43181 rows × 4 columns
```

Figure 1: We can see that the districts on both sides(children and household) are matching

3 METADATA OF RECODED VARIABLES

Along with this document, an Excel file is also attached (`https://docs.google.com/spreadsheets/d/1ypyREMqNoYry7VJ2rSggHw-_NGCYxtGLv-0UZNYyAtI/edit#gid=0`). Use the link above for a better viewing experience. This contains all the variables that have been recoded. Information such as the variables used to get the recoded variable, definition, stats, etc is included.

References for where the variables have been taken from have been mentioned in the definition column, in case the reference is not mentioned, you can assume the variables have been created following the guidelines from Mira.

An example of how to read the metadata for the table below:

New Variable	Definition	Variables Used	Values	Stats
`zero_dose`	Not received any doses of DPT or Penta	h3,h51	h3!=(1,2,3) or h51!=(1,2,3), T h3==(8,9) or h51==(8,9), NULL else, F	Counts of the True, False and missing values

The variable '`zero_dose`' was created using variables h3,h51. The definition of what zero dose is provided in the definition column. The value that '`zero_dose`' has been recoded into is mentioned in the values column. Stats on the number of True, False and missing values are present in the other columns.

Verifying if the sum of vaccination outcomes is correct:

False	True	Missing	Total
Zero_dose 43181	38196	4985	0
Fully_vac 43181	9226	33507	448
Under_vac 43181	33507	9226	448
Unimmunised 43181	41460	1721	0
Purna_tika 43181	43181	0	0

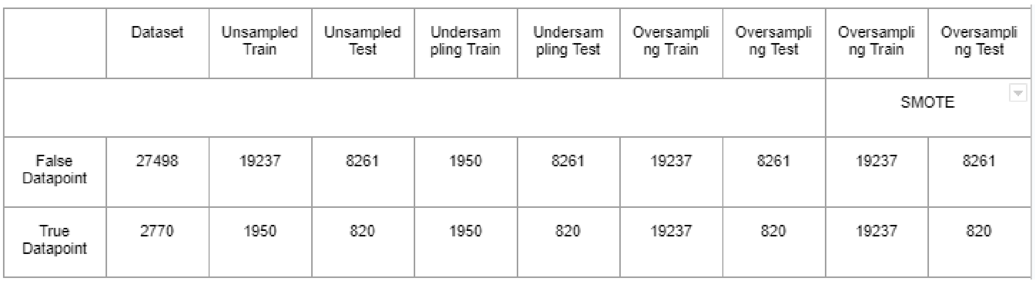
4 MODELING APPROACH

I have attached a notebook file on a logistic regression modeling approach with this document. Using the flag and setting it to 1, we get a dataset consisting of only binary values. This is taken as input for the script and modeled on this dataset. Certain variables are removed as they have excessive empty values. Variables used in modeling are provided in the appendix.
The modelling approach used here is logistic regression, and in order to run such a model, the best input is binary input and not continuos input. Hence, I converted a few variables into binary variables. Let's look at the dataset at first. We have a total of 43,181 datapoints. There are many empty values in this, so we remove all rows containing any empty values and end up with 31,457 datapoints. Below is the percentage of datapoints of True values of `zero_dose` kids.

All datapoints	After dropping empty values
4985	2640
11.54 %	8.77 %

The distribution of the values for `zero_dose` is 27,498 values as False and 2,770 values as True. We see an imbalance in the ratio of True and False values. To counter this, we will use a few undersampling and oversampling approaches.

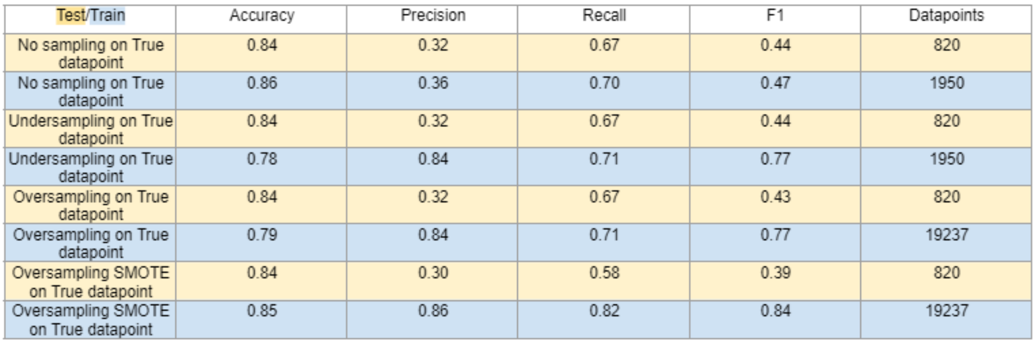
Below is a table with the counts after using a few sampling approaches. The samplings have been achieved using imblearn library.



	Dataset	Unsampled Train	Unsampled Test	Undersampling Train	Undersampling Test	Oversampling Train	Oversampling Test	Oversampling Train	Oversampling Test
								SMOTE	
False Datapoint	27498	19237	8261	1950	8261	19237	8261	19237	8261
True Datapoint	2770	1950	820	1950	820	19237	820	19237	820

Figure 2: Datapoints count

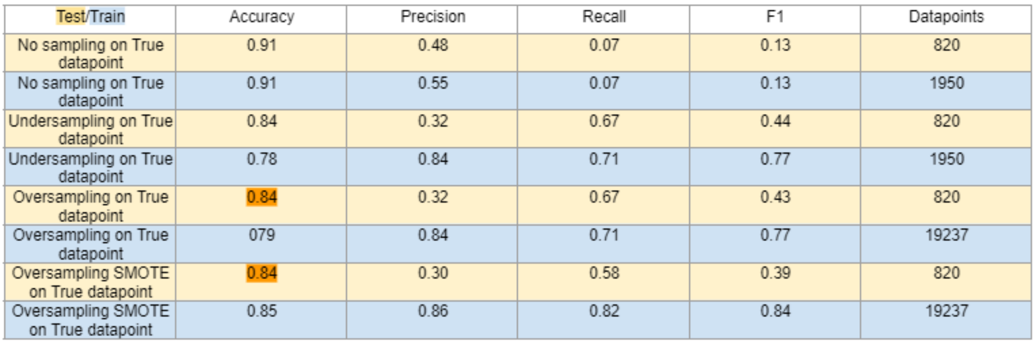
Using this, a model was trained on logistic regression. Details of it's accuracy is given below.



Test/Train	Accuracy	Precision	Recall	F1	Datapoints
No sampling on True datapoint	0.84	0.32	0.67	0.44	820
No sampling on True datapoint	0.86	0.36	0.70	0.47	1950
Undersampling on True datapoint	0.84	0.32	0.67	0.44	820
Undersampling on True datapoint	0.78	0.84	0.71	0.77	1950
Oversampling on True datapoint	0.84	0.32	0.67	0.43	820
Oversampling on True datapoint	0.79	0.84	0.71	0.77	19237
Oversampling SMOTE on True datapoint	0.84	0.30	0.58	0.39	820
Oversampling SMOTE on True datapoint	0.85	0.86	0.82	0.84	19237

Figure 3: Unbalanced accuracy

After going through a few parameters in logistic regression, i came across a parameter `class_weight` in the logistic regression model. This would assign the minority class a higher weight, hence I tried using this and compared the results with my previous approach.



Test/Train	Accuracy	Precision	Recall	F1	Datapoints
No sampling on True datapoint	0.91	0.48	0.07	0.13	820
No sampling on True datapoint	0.91	0.55	0.07	0.13	1950
Undersampling on True datapoint	0.84	0.32	0.67	0.44	820
Undersampling on True datapoint	0.78	0.84	0.71	0.77	1950
Oversampling on True datapoint	0.84	0.32	0.67	0.43	820
Oversampling on True datapoint	079	0.84	0.71	0.77	19237
Oversampling SMOTE on True datapoint	0.84	0.30	0.58	0.39	820
Oversampling SMOTE on True datapoint	0.85	0.86	0.82	0.84	19237

Figure 4: Balanced accuracy

I noticed the accuracy to be slightly higher in the latter case.

In order to find out with variables must be included in the analysis, we did run a pearson and cramer V correlation. It did not assist us much as the correlations weren't strong.

5 FUTURE WORKS

Currently, the variables used for modeling are an exhaustive list taken through recommendations. In the next modeling stage, better variable clusters can be considered with correlation and better outcomes.
Also, modeling has been currently done on logistic regression. Other modeling methods can be used such as Bayesian modeling which can give better results for such datasets.

A Appendix for added variables

Variables used from JMP (Joint Monitoring Programme (JMP) Methodology. 2017.) are listed below
impr_ws,unimpr_ws,basic_drinking_w,limited_drinking_w,jmp_w8,jmp_w2,jmp_w5,jmp_s1,jmp_s6,j

Variables used in modeling are listed below:
’impr_ws’, ’unimpr_ws’, ’basic_drinking_w’, ’limited_drinking_w’, ’jmp_w8’, ’jmp_w2’, ’jmp_w5’, ’jmp_s1’, ’jmp_s6’, ’jmp_s8’, ’jmp_s7’, ’highest_grade_comp’, ’jmp_h1’, ’jmp_h2’, ’jmp_h3’, ’wi_combined_poor’, ’wi_ur_poor’, ’wi_statewise_poor’, ’wi_statewise_ur_rs_poor’, ’electricity’, ’kaccha_floor’, ’kaccha_roof’, ’kaccha_walls’, ’all_kaccha_house’, ’own_house’, ’own_agri_land’, ’bpl_card’, ’insurance’, ’clean_fuel_usage’, ’caste_General’, ’caste_OBC’, ’caste_SC’, ’caste_ST’, ’bank_acc’, ’highest_edu_lvl_Higher’, ’highest_edu_lvl_No education’, ’highest_edu_lvl_Primary’, ’highest_edu_lvl_Secondary’, ’w_religion_Buddhist / Neo_Buddhist’, ’w_religion_Christian’, ’w_religion_Hindu’, ’w_religion_Jain’, ’w_religion_Muslim’, ’w_religion_No religion’, ’w_religion_Parsi / Zoroastrian’, ’w_religion_Sikh’, ’child_death’, ’w_marital_status_Married’, ’w_marital_status_Never in union/marriage’, ’w_marital_status_widowed divorced separated deserted’, ’mcp_card’, ’antenatal_care’, ’antenatal_4plus’, ’tetanus’, ’birth_personnel’, ’delivery_place_Home’, ’delivery_place_Private’, ’delivery_place_Public’, ’delivery_financial_assistance’, ’delivery_jsy’, ’baby_checkup_2mnts’, ’modern_contraceptive’, ’icds_rec’, ’icds_rec_bf’, ’any_anaemia’, ’preg_wm_any_anem’, ’union_before_15’, ’union_before_18’, ’owns_phone’, ’internet’, ’stunting’, ’stunting_severe’, ’wasting’, ’wasting_severe’, ’underweight’, ’underweight_severe’, ’zero_dose’, ’fully_vac’, ’under_vacc’, ’unimmunised’,’purna_tika’

Variables when Categorical_eleminate_flag is set to 1:
’impr_ws’, ’unimpr_ws’, ’basic_drinking_w’, ’limited_drinking_w’, ’jmp_w8’, ’jmp_w2’, ’jmp_w5’, ’jmp_s1’, ’jmp_s6’, ’jmp_s8’, ’jmp_s7’, ’highest_grade_comp’, ’jmp_h1’, ’jmp_h2’, ’jmp_h3’, ’wi_combined_poor’, ’wi_ur_poor’, ’wi_statewise_poor’, ’wi_statewise_ur_rs_poor’, ’electricity’, ’kaccha_floor’, ’kaccha_roof’, ’kaccha_walls’, ’all_kaccha_house’, ’own_house’, ’own_agri_land’, ’bpl_card’, ’insurance’, ’clean_fuel_usage’, ’caste_General’, ’caste_OBC’, ’caste_SC’, ’caste_ST’, ’bank_acc’, ’highest_edu_lvl_Higher’, ’highest_edu_lvl_No education’, ’highest_edu_lvl_Primary’, ’highest_edu_lvl_Secondary’, ’w_religion_Buddhist / Neo_Buddhist’, ’w_religion_Christian’, ’w_religion_Hindu’, ’w_religion_Jain’, ’w_religion_Muslim’, ’w_religion_No religion’, ’w_religion_Parsi / Zoroastrian’, ’w_religion_Sikh’, ’child_death’, ’w_marital_status_Married’, ’w_marital_status_Never in union/marriage’, ’w_marital_status_widowed divorced separated deserted’, ’residing_husband’, ’other_wives’, ’mcp_card’, ’antenatal_care’, ’antenatal_4plus’, ’tetanus’, ’birth_personnel’, ’delivery_place_Home’, ’delivery_place_Private’, ’delivery_place_Public’, ’delivery_financial_assistance’,

'delivery_jsy', 'baby_checkup_2mnts', 'modern_contraceptive', 'icds_rec', 'icds_rec_bf', 'any_anaemia', 'preg_wm_any_anem', 'union_before_15', 'union_before_18', 'owns_phone', 'internet', 'stunting', 'stunting_severe', 'wasting', 'wasting_severe', 'underweight', 'underweight_severe', 'zero_dose', 'fully_vac', 'under_vacc', 'bcg', 'bcg_card', 'polio_0', 'polio_0_card', 'dpt_1', 'dpt_1_card', 'pentavalent_1', 'pentavalent_1_card', 'hepatitis_b', 'hepatitis_b_card', 'rotavirus_1', 'rotavirus_1_card', 'dpt_1_booster'

Variables when Categorical_eleminate_flag is set to 0:
'impr_ws', 'unimpr_ws', 'basic_drinking_w', 'limited_drinking_w', 'jmp_w8', 'jmp_w2', 'jmp_w5', 'jmp_s1', 'jmp_s6', 'jmp_s8', 'jmp_s7', 'jmp_h1', 'jmp_h2', 'jmp_h3', 'wi_combined', 'wi_combined_score', 'wi_ur', 'wi_ur_score', 'wi_statewise', 'wi_statewise_ur_rs', 'wealth_q_HV270', 'wealth_q_HV271', 'electricity', 'kaccha_floor', 'kaccha_roof', 'kaccha_walls', 'all_kaccha_house', 'own_house', 'own_agri_land', 'bpl_card', 'insurance', 'clean_fuel_usage', 'caste', 'highest_grade_comp', 'bank_acc', 'highest_edu_lvl', 'highest_edu_year', 'w_religion', 'child_death', 'child_sex_ratio_statewise', 'child_sex_ratio_districtwise', 'child_sex_ratio_clusterwise', 'w_marital_status', 'residing_husband', 'other_wives', 'mcp_card', 'antenatal_care', 'antenatal_4plus', 'tetanus', 'birth_personnel', 'delivery_place', 'delivery_financial_assistance', 'delivery_jsy', 'baby_checkup_2mnts', 'modern_contraceptive', 'icds_rec', 'icds_rec_bf', 'w_age_marr', 'any_anaemia', 'preg_wm_any_anem', 'union_before_15', 'union_before_18', 'owns_phone', 'internet', 'stunting', 'stunting_severe', 'wasting', 'wasting_severe', 'underweight', 'underweight_severe', 'zero_dose', 'fully_vac', 'under_vacc', 'bcg', 'bcg_card', 'polio_0', 'polio_0_card', 'polio_doses', 'dpt_1', 'dpt_1_card', 'dpt_doses', 'pentavalent_1', 'pentavalent_1_card', 'pentavalent_doses', 'hepatitis_b', 'hepatitis_b_card', 'hepatitis_b_doses', 'rotavirus_1', 'rotavirus_1_card', 'rotavirus_doses', 'je_doses', 'measles_doses', 'dpt_1_booster'