Household poverty classifier

Artifoni Mattia 807466
Bottoni Federico 806944
Capra Riccardo 808227

- Problem of poverty in Costa Rica
- Improvement of Proxy Means Test (or PMT)
- Improvement of the submitted results

https://www.kaggle.com/c/costa-rican-household-poverty-prediction

Dataset and data manipulation

Raw dataset: features

Dataset rows describe the Costa Rica residents and their features related to the welfare level

The columns are 142 and they define the condition of the subject in the following parameters.

- Members of the household
- Material and structural condition of the house
- Toilet, electricity system and rubbish disposal
- Role of the person in the family
- Education
- Economic situation
- Region where he/she lives

Raw dataset: labels

The labels are the 4 levels of welfare encoded in the target field

Code	Description	Freq.	Rel.Freq.
1	*extreme poverty*	755	7.89%
2	*moderate poverty*	1597	16.71%
3	*vulnerable households*	1209	12.65%
4	*non-vulnerable households*	5996	62.74%

Processed dataset: features

Dataset rows and columns were deleted/fixed accordingly to the task needs

The columns resulted from the manipulation steps are 106.

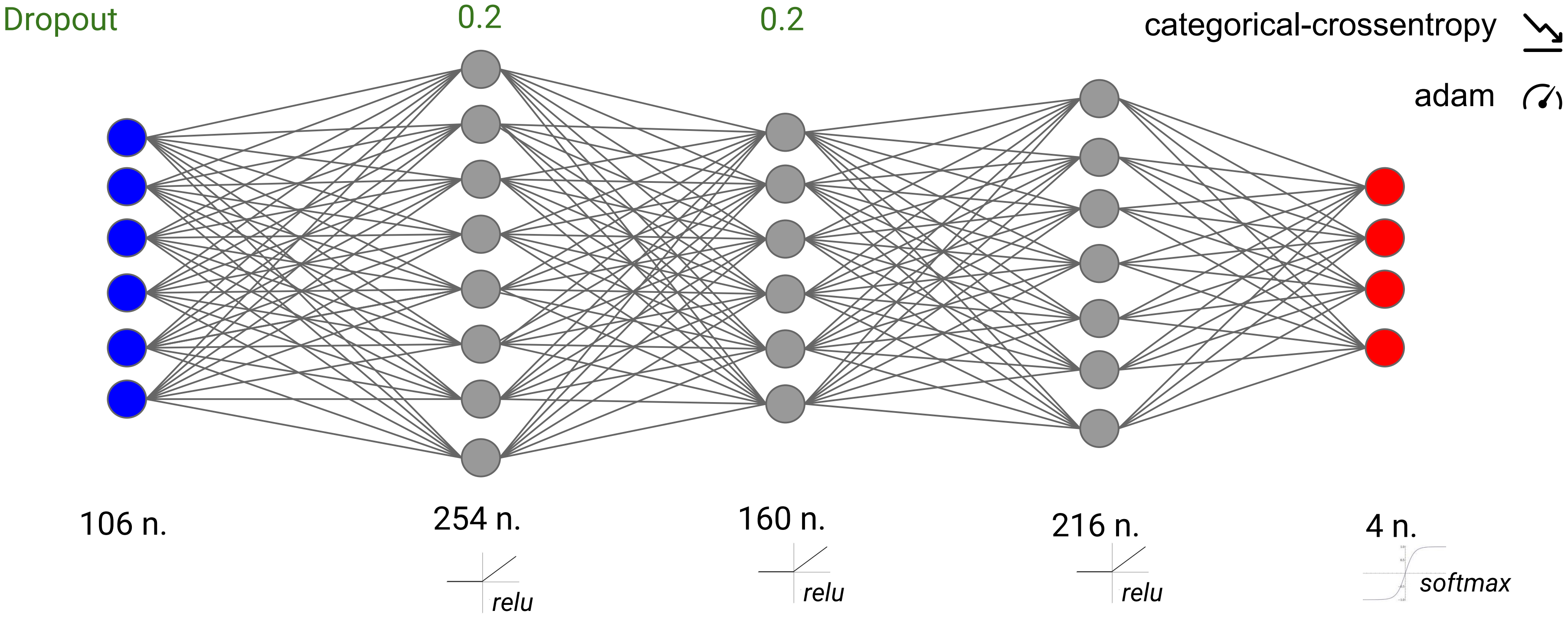
- Deletion of the redundant or useless columns for the presented task
- Statistical filling or simple deletion of the incomplete rows/cells
- Manipulation of all the values to the float format (e.g. yes->1, .09->0.09)
- Rounding of periodic numbers to 2 decimal values

Base model and AutoML

Data preprocess

- Standardized data
- Oversampling
- Shuffle
- Split 0.9 - 0.1
- Categorical label

Deep neural network



Dropout 0.2 0.2

categorical-crossentropy

adam

106 n.

254 n. relu

160 n. relu

216 n. relu

4 n. softmax

Hyperparameters sampling

Gaussian Process

Acquisition: ExpectedImprovement
Covariance: SquaredExponential
Evaluation: 10-fold-crossvalidation

20 iterations

Random Forest

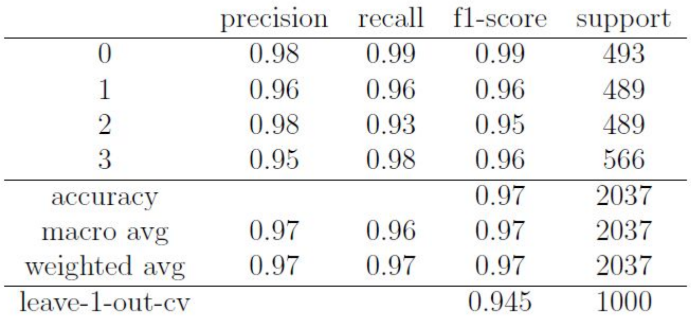
Acquisition: ExpectedImprovement
Evaluation: 10-fold-crossvalidation

20 iterations

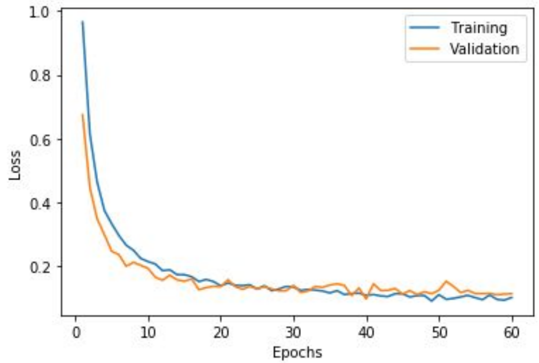
HP	Range	Sampled value
learning-rate	[0.001, 0.01]	0.002090922710
beta1	[0.8, 0.999]	0.919947158421
beta2	[0.8, 0.999]	0.978863157785
hidden.n.1	[8, 256]	254
hidden.n.2	[8, 256]	160
hidden.n.3	[8, 256]	216

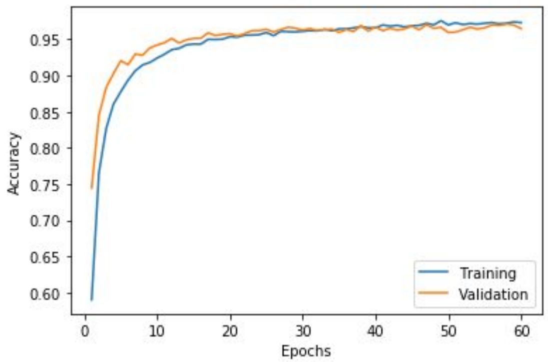
Performances evaluation

- Based on 10% of the dataset
- Leave1Out on 1000 records (4.9%)



	precision	recall	f1-score	support
0	0.98	0.99	0.99	493
1	0.96	0.96	0.96	489
2	0.98	0.93	0.95	489
3	0.95	0.98	0.96	566
accuracy			0.97	2037
macro avg	0.97	0.96	0.97	2037
weighted avg	0.97	0.97	0.97	2037
leave-1-out-cv			0.945	1000



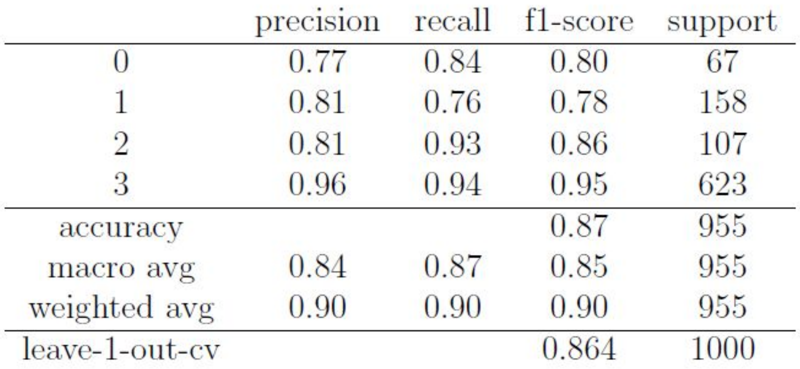


Custom tests

Without oversampling

Remotion of the oversampling step from data preprocessing

- Training is faster
- Each metric is more than 10% lower
- Class 3 metrics (non-vulnerable households) are similar to the base

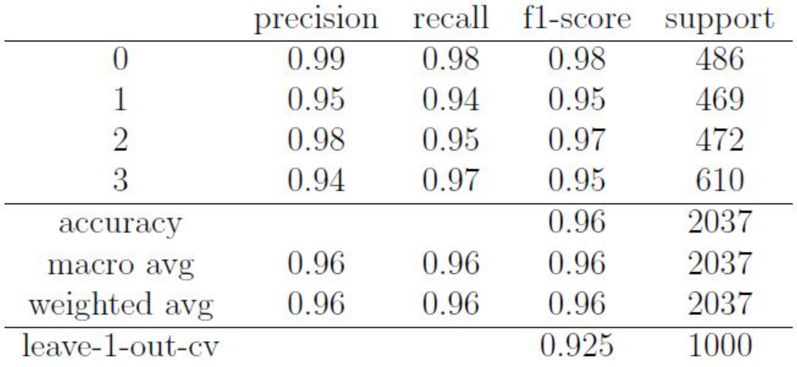


	precision	recall	f1-score	support
0	0.77	0.84	0.80	67
1	0.81	0.76	0.78	158
2	0.81	0.93	0.86	107
3	0.96	0.94	0.95	623
accuracy			0.87	955
macro avg	0.84	0.87	0.85	955
weighted avg	0.90	0.90	0.90	955
leave-1-out-cv			0.864	1000

Grouping of data

The columns that refer to a same feature of the person were grouped according to the logic

- Features from 106 to 93
- Grouping process is lossless
- All metrics are really similar (slightly lower)
- Training is faster



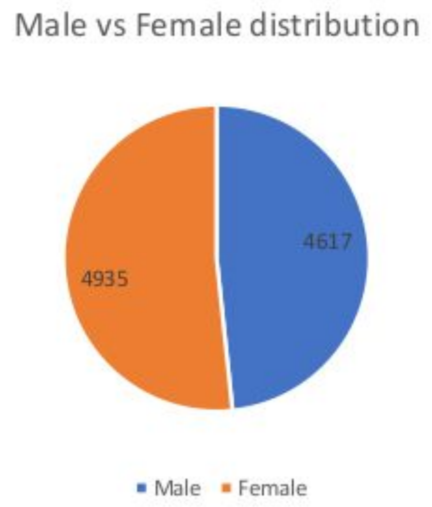
	precision	recall	f1-score	support
0	0.99	0.98	0.98	486
1	0.95	0.94	0.95	469
2	0.98	0.95	0.97	472
3	0.94	0.97	0.95	610
accuracy			0.96	2037
macro avg	0.96	0.96	0.96	2037
weighted avg	0.96	0.96	0.96	2037
leave-1-out-cv			0.925	1000

Fairness study

Gender distribution

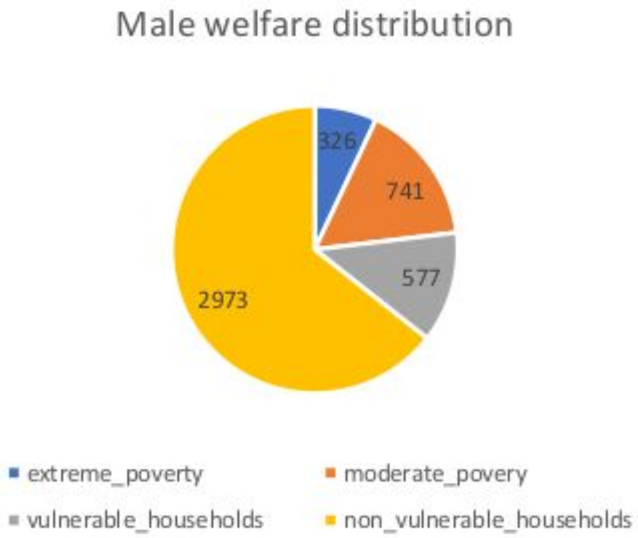
Study of the distribution of males and females

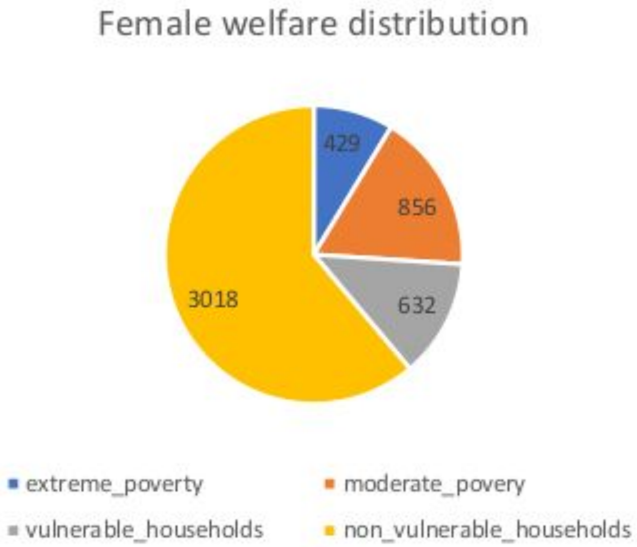
- The distribution of males and females entities in the dataset was balanced
- 51.7% females and 48.3% males



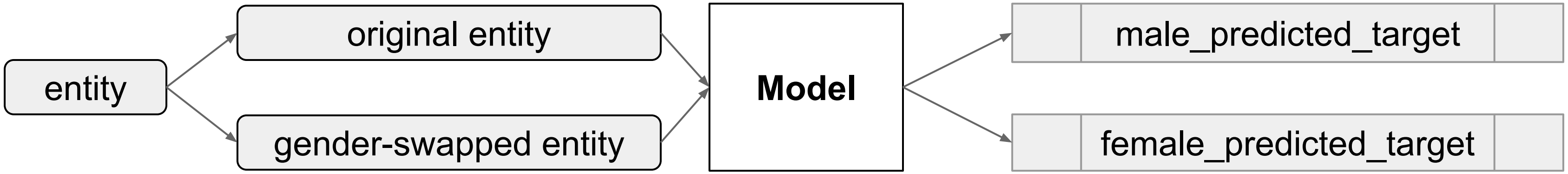
Wealth distribution

- Cosine similarity between the two distribution: 0.9990


Male welfare distribution


Female welfare distribution

Model fairness


entity

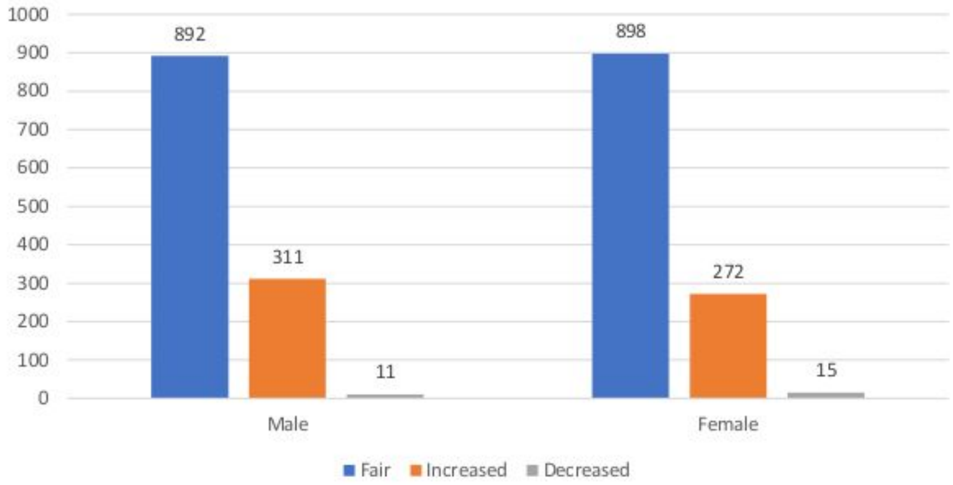
original entity

gender-swapped entity

Model

male_predicted_target

female_predicted_target



Conclusions

- Dataset unbalancing and dimensions
- Good results
- Limits and uses of the model

https://github.com/federicobottoni/household-poverty-classifier