Random Forest and the Effects of the Lockdown on Household Income - Evidence From NIDS-CRAM Wave 1

Johannes Coetsee - 19491050[a]

[a] *Stellenbosch University*

Keywords:

JEL classification

1. Introduction

The purpose of this paper is to report on the implementation of a Random Forest (RF) algorithm for a classification-type problem, namely, to classify which individuals and households were more likely to lose their main source of income due to the coronavirus and subsequent lockdown in South Africa in March and April 2020.

RF is well-suited for these types of problems

Data

This study utilises the first wave of the National Income Dynamics Study - Coronavirus Rapid Mobile Survey 2020 (NIDS-CRAM) dataset, a longitudinal telephonic household survey conducted by the Southern Africa Labour and Development Research Unit (SALDRU) in April and May 2020. NIDS-CRAM investigates the various social and economic effects of the national lockdown implemented in March 2020, and more broadly, the consequences of the global pandemic on the South African context.

In total, the dataset consists of 21 features, which is reported in Table **??** below, with 6838 observations for each feature. The main variable of interest is 'Income.Change' - a binary variable where a value of 1 indicates that the household has lost their main source of income, whilst 2 indicates that it has not. The question asked to respondents reads as : "Has your household lost its main source of income since the lockdown started on 27th March?"

Email address: 19491050@sun.ac.za - https://github.com/Coetsee (Johannes Coetsee - 19491050)

1.1. Missing Values and Transformations

Some missing values are imputed using simple median-replacement, where the missing value is replaced with the median value computed on the rest of the data. This method is preferred for continuous variables over other data types.

2. Methodology

2.1. random forest

2.2. sampling

2.3. bias variance trade-off

2.4. prediction and confusion matrix, train vs test data

2.5. error rate and bootstrap samples

2.6. number of nodes

2.7. hyperparameter tuning

2.8. variable importance

2.9. partial dependence plot

3. Results

4. Conclusion

References

Appendix

Appendix A