Utilising Random Forest Algorithms to Classify Those Most Likely to Lose Their Main Source of Income due to Lockdown- Evidence From NIDS-CRAM Wave 1

Johannes Coetsee - 19491050[a]

[a]*Stellenbosch University*

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.[1] RF is well-suited for classification-type problems, as

2. 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 population.

In total, the dataset consists of 21 features, which is reported in Table **??** below, with 7073 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]The template for this report is based on that provided by Katzke (2017).

Missing Values and Transformations

In order to avoid losing information, this study imputes missing values for the entire dataset of 21 features. Before imputation, however, NAs for the 'Employment.Type' feature are replaced by 0's to indicate 'unemployed', as this survey question was only asked to those who were employed. Those who refused to respond or did not know their main form of work, were indicated as missing and therefore imputed. Similarly, system NAs for the 'Tertiary' feature - a dummy variable indicating whether an individual completed some form of tertiary education - was replaced by 0, or 'no', as this question was only asked to those who were eligible. Although not perfect solutions, these are fair assumptions to make in order to include these potentially important variables. Additionally, the feature indicating in which Municipal Demarcations Board District Council the household is situated was transformed into a matrix of binary variables so as to accommodate the necessary structure needed for imputation.[2]

The method of imputation used in this paper draws on a random forest algorithm to impute missing values trained on the matrix of observed values in the data. This can be done using the package *missForest* in R, following a two-step procedure. First, missing values are pre-imputed using simple median replacement - where the missing value is replaced with the median value computed on the rest of the observed data for each continuous feature. For categorical variables, missing values are replaced by the most frequently occurring non-missing value.[3] Second, a forest is grown using multivariate splitting, where the splitting rule is averaged only over non-missing values. Data is then imputed by regressing each feature on all other features, thereafter predicting missing values using the fitted forest. This process is iterated in order to update the initial median-replaced values until the stopping criterion - in our case, when the difference between the previous iteration and new iteration have become larger once for each data type - is met (Tang & Ishwaran, 2017).

The usage of this algorithm is necessitated by the nature of the data, where features are of three different data types, namely, categorical, numeric and continuous. Stekhoven & Bühlmann (2012) show that this iterative imputation procedure outperforms many other widely-used implementation methods such as, for instance, K-Nearest Neighbours (KNN) imputation and Multivariate Imputation by Chained Equations (MICE), especially within mixed-type data contexts. Furthermore, it inherits all the positive attributes attributed to random forests itself, such as being robust to noisy data due to inherent feature selection, as well as being simple to implement. However, it is computationally intensive and crucially also relies on the assumption that missing data are Missing At Random (MAR). If not MAR, there is possibility of introduced selection bias. This is deemed a permissible admission due to the relatively low number of missing values in the dataset as a whole.

[2]This is due to the fact that only 53 levels are allowed for factored variables using the *missForest* and *randomForest* packages, whereas the District Council variable consists of 54.
[3]This process is also called Strawman imputation.

3. Methodology

Cloud Computing

SQL

The Random Forest Algorithm

GBM

4. Results

4.1. Model 1: Random Forest

prediction and confusion matrix, train vs test data

error rate and bootstrap samples

number of nodes

hyperparameter tuning

variable importance

partial dependence plot

4.2. Model 2: GBM Random Forest

gri

5. Conclusion

References

Katzke, N.F. 2017. *Texevier: Package to create elsevier templates for rmarkdown*. ed. Stellenbosch, South Africa: Bureau for Economic Research.

Stekhoven, D.J. & Bühlmann, P. 2012. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 28(1):112–118.

Tang, F. & Ishwaran, H. 2017. Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*. 10(6):363–377.

Appendix

Appendix A