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 Ingle *et al.* (2020).[2]

In total, the dataset consists of 21 features, reported in Table 2.1 below, with 7073 observations in total. Table 2.1 also reports the amount of missing values for each feature, as well as a relevant description.[3] The main feature 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 follows: "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).

[2] The data is publicly available at https://www.datafirst.uct.ac.za/.

[3] In this case, the survey answers 'Refused', 'Don't Know', Not Applicable and 'Missing' are all defined as NAs.

Selected Features	NAs	Description
Income Change	168	Has household lost main source of income since lockdown
Sources Income Decreased	376	Did sources of household income decrease during lockdown
Employed	161	Employment Status
Employment Type	151	Respondent's main form of work (0 = unemployed)
Sources HH Income	121	Sources of household income in February
Children Change	93	Change in number of children in house compared to pre-lockdown
Province	8	Province currently living in now
Dwelling Type	6	Type of dwelling, whether house, informal, traditional or other
Race	0	Respondent's given population group
Geo Type	8	Geography Type (derived from 2011 census)
HH Income Apr	2665	Total household income after tax in April
Moved	8	Whether respondent moved to another dwelling for lockdown
Grant	36	Whether the respondent receives any kind of government grant
Electricity Access	2	Whether dwelling has access to electricity
Water Access	5	Whether dwelling has piped or tap water
HH Size	32	Number of people resident (Household Size)
Education	45	Highest school grade completed
Tertiary	9	Has respondent successfully completed some tertiary education
Age	0	Respondent's age in years
Age Interval	0	Age interval (5 year intervals)
Gender	0	Respondent's stated gender
District Council	8	Municipal Demarcations Board District Council (from 2011 Census)

Table 2.1: Features

Missing Values and Transformations

It is evident from 2.1 that missing values might be a stumbling block for accurate analyses using this data. In particular, the 'HH Income Apr' variable has a large amount of NAs, most of which are attributed to the 'Don't Know' category on the questionnaire. In other words, respondents reported that they did not know their exact level of income for the month of April 2020. In order to avoid losing information, this study imputes these missing values as well as for all other features within the dataset. Furthermore, 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 meaningful variables. Additionally, the feature indicating in which District Council the household is situated was transformed into a matrix of binary variables so as to accommodate the necessary structure needed for imputation.[4]

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, which follows 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 the categorical data type, missing values are replaced by the most frequently occurring non-missing value.[5] 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 specific 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) and Tang & Ishwaran (2017) show that this iterative RF 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 characteristics 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.[6] This is deemed a permissible admission due to the relatively low number of missing values in the dataset as a whole. For the most problematic feature, 'HH Income Apr', the imputation strategy above is especially relevant as missing values are more likely to be those closer to the median-income group than to, for instance, the mode or mean incomes.

3. Methodology

[4] This is due to the fact that only 53 levels are allowed for factored variables using both the *missForest* and *randomForest* packages, whereas the District Council variable consists of 54 levels.

[5] This process is also called Strawman imputation.

[6] missForest is not unique in requiring this assumption, however.

Computing

All computation was done using the Amazon Web Services' (AWS) Elastic Compute Cloud (EC2) service, combined with the functionality of RStudio Server. A 't2.micro' virtual machine instance was created with a public IP address, through which RStudio Server was initiated.

Computing time -

Cloud computing is necessitated by especially the grid-searches employed in later sections due to computing limitations on the local machine, however, clustering was not deemed necessary due to the relatively small size of the data.

Furthermore, parallel programming was utilised for the more computationally intensive tasks such as missing value imputation and parameter tuning (using the Parallel sockets approach for windows).

SQL

SQL was used in two ways in this study. First, *sqlite3* databases were created for the separate tables entitled 'nids' and 'derived', and subsequently compiled in one large database containing all features of the raw data. This was done using SQL syntax and the function *dbConnect* within RStudio. However, using the Bash Unix shell for Windows[7], these tables were queried and data was surveyed in order to select the relevant features necessary for the RF implementation. Although not necessary to use SQL to this end, it is more efficient in terms of memory usage than reading the larger datasets directly into R's memory. After feature selection, the final dataset to be read into R was once again compiled and collected within RStudio using SQL syntax.

Algorithms

This study compares the performance of three different classification algorithms: 1) a random forest, 2) a simple Gradient Boosted Random Forest, and 3)

RF

GBM

4. Results

[7] Made accessible due to being a member of the Windows Insider Program.

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

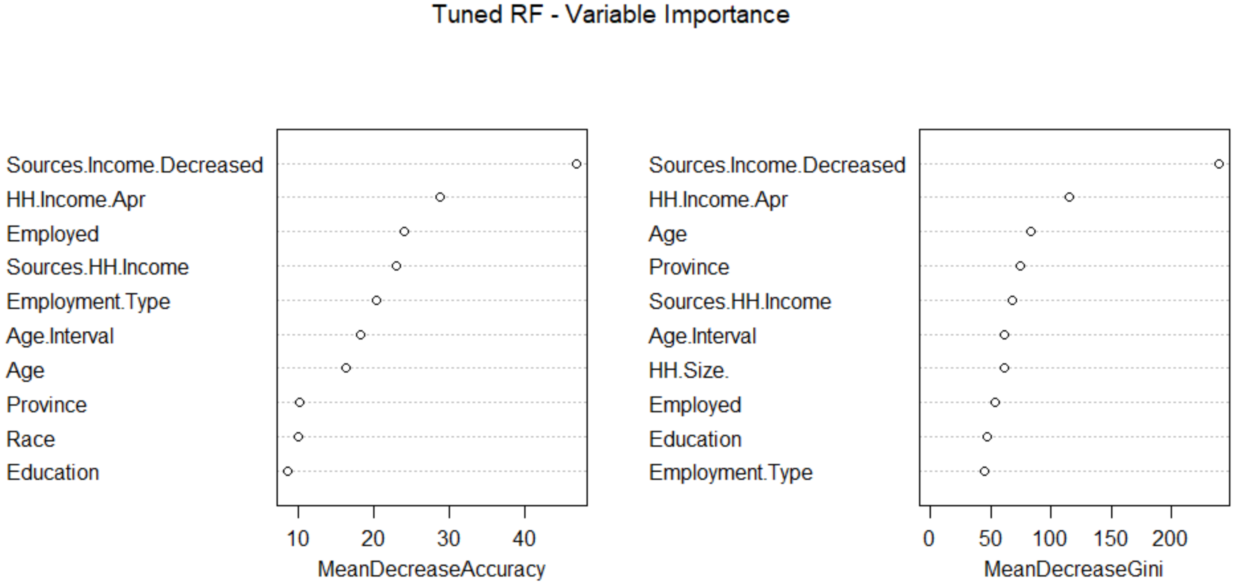
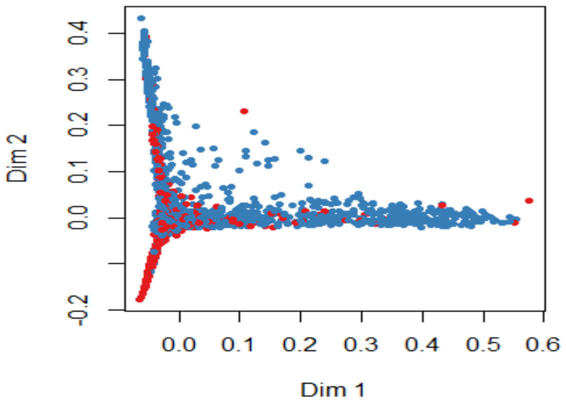


Figure 4.1: - RF Tuned Model

4.1.1. MDS plots



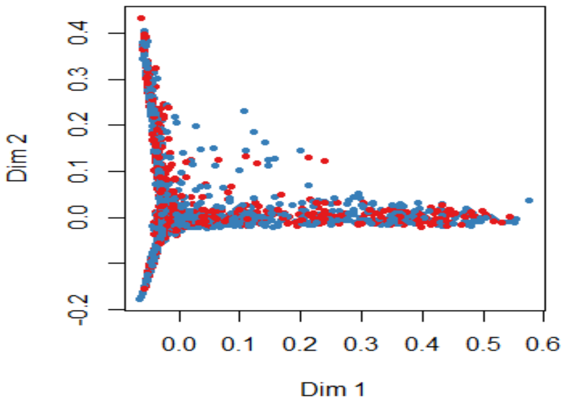


Figure 4.2: - Training Data (Top), Test Data (Bottom)

4.2. Model 2: GBM Random Forest

gri

5. Conclusion

References

Ingle, K., Brophy, T. & Daniels, R. 2020. National income dynamics study–coronavirus rapid mobile survey (nids-cram) panel user manual. *Technical Note Version*. 1.

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