Research Proposal

Siphiwe Bogatsu

2024-03-10

Introduction

Youth people in South Africa continue to be unable to find work and withdrawing from the labour force entirely, with youth unemployment and underemployment overwhelmingly standing at a rate above that of the national average. According to StatSA (2024), for the first quarter of 2022 the unemployment rate was 63,9% for those aged 15-24 and 42,1% for those aged 25-34 years, while the current official national rate stands at 34,5%. Being able to build models of reality that can predict, and understand, which young people will find employment and which ones will require additional help, helps promote evidence-based decision-making, supports economic empowerment, and allows young people to thrive in their chosen careers. However, youth survey datasets present - useful in building these models - are high dimensional in their nature. This presents a critical challenge of how to reduce the dimension of these datasets and the ability to extract relevant features as means towards building robust models to predict which young people will find employment.

Principal component analysis (PCA) has been widely used approach in data dimensionality reduction and data processing for use in fitting machine learning models (Zou et al 2006). Even though this approach has led to fruitful stories in numerous applications in biology, engineering, and social science, it has an obvious drawback. That is, PCA reduces data into lower dimensional space called the principal components (PCs), to maintain the same structure as the original data. Essentially, Each PC is weighted average of all the predictors with a loading coefficient assigned to each predictor. Each PC uses all the variables regardless if a predictor is a noise - meaning no true effect in the PC. This raises serious concerns, because would not it better to exclude the predictor during the construction of PCs such that they become more robust ? The result of using all predictors, including the noisy ones, will generate PCs likely contaminated with noise such that the resulting PCs may deviate significantly from the original data (Hsu, Huang & Chen, 2014). Another drawback is that when all the coefficient loading are nonzero for each PCs, as evidenced by the standard PCA, it is quite difficult to interpret the derived PCs (Zou et al, 2006).

Drawing from Hsu, Huang & Chen (2014)'s cancer research paper, I propose applying sparse principal component analysis (SPCA) to help with data dimension and select important variables simultaneously on youth survey dataset. Therefore, I am interested in evaluating the sparse PCA's effectiveness, as an upgrade from PCA, to reduce the dimensionality of youth survey data and subsequently exclude ineffective predictors from the PCA model.

Data Description

To illustrate the potential application of sparse PCA, youth survey data of 2022-2023 of 4020 South African youth is used for demonstration. Demographic information includes gender (female: 2269, male: 1751);their mean age of 1997. Geographical information include geography (urban:2797, rural: 803, suburb:420).

The data was collected by a company situated in Stellenbosch, Predictive Insights, through 4 rounds of a survey of youth in the South African labour market, conducted at 6-month intervals. Each youth in the dataset was surveyed one year prior to the follow-up survey, which specifically asked whether the youth is currently employed.

Analysis Approach

The response variable is of binary nature (0: unemployed and 1: employed) describing whether the youth is employed after a year since the baseline survey. Below, over 72% of youth are unemployed after a year.

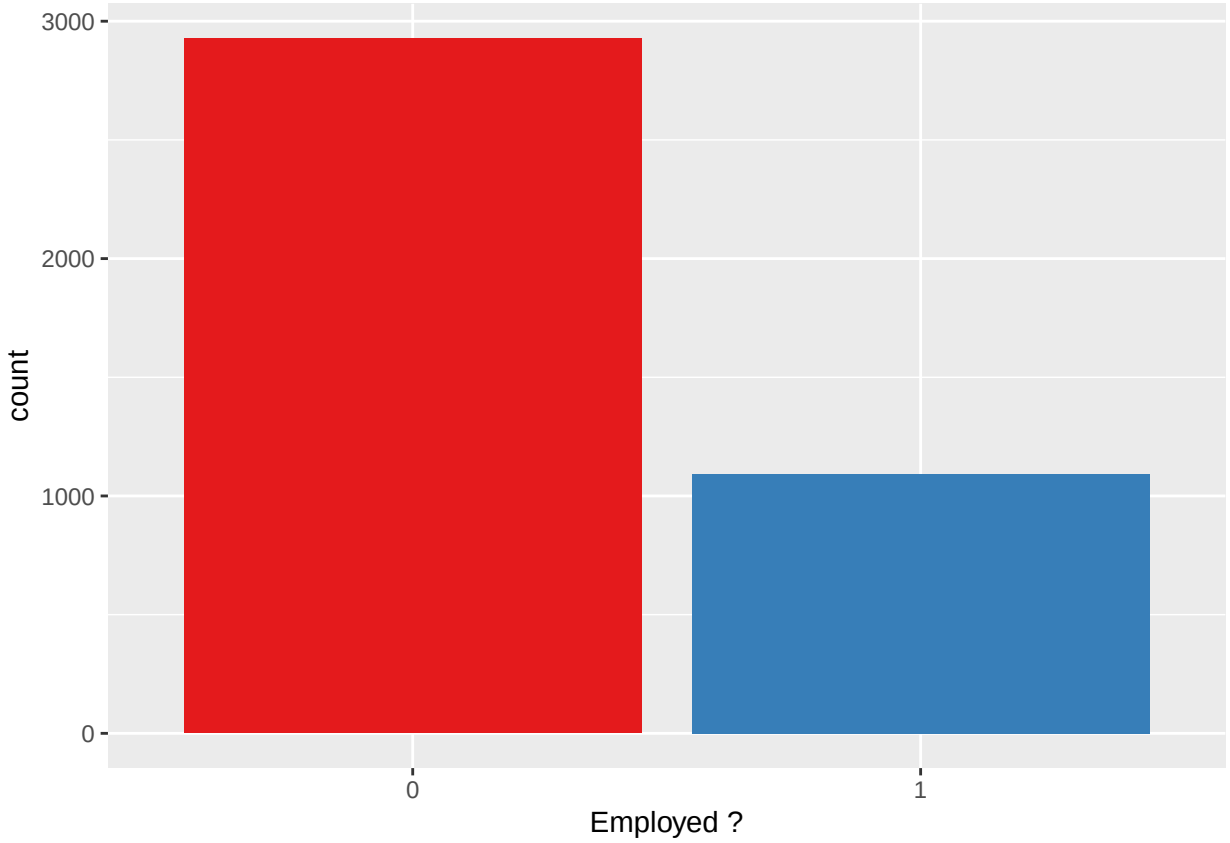

Table 1: Predictors used in this study.

Variable	Description	Data Type
Round	The four rounds of 6-month interval survey.	numeric
Status	Labour market status of individual	character
Tenure		numeric
Geography	rural, suburb, or urban ?	character
Province	provincial residence of youth	character
Matric	Whether or not the youth has a matric certificate.	numeric
Degree	Whether or not the youth has a degree	numeric
Diploma	Whether or not the youth has a diploma	numeric
School Quintile	South Africa's education system is divided into five quintiles based on the socio-economic status of the area surrounding the school	numeric
Math	Mathematics results in matric level mark	numeric
Mathlit	Mathematical Literacy in matric level mark	numeric
Additional_lang	Additional language in matric level mark	numeric
Home_lang	Home language in matric level mark	numeric
Science	Physical Science or Life Sciences in matric level mark	numeric
Female	Gender of youth	numeric
Age	Age of youth	numeric

Variable	Description	Data Type
Birth Month	Birth month of youth in the calendar	numeric

For this study, I will use XG Boosting model to fit the predictors on the target variable. Briefly, stochastic gradient boosting operates on a premise that ensemble many weak learners into strong learners by letting the base models to sequentially learn from the mistakes made by the previous ones. In this case, there are three main hyperparameters to tune: number of trees, shrinkage parameter and the interaction depth.