Computing Lab / Data Warehousing and Business Intelligence

Analytics on the US Nonprofit Sector - Milestone V

Because the nonprofit sector is a growing trillion-dollar sector, we designed our analysis so that our dataset would generate insight around non-profit profit-making. Thus we selected total revenue as our target variable and devised a method to select predictor values to build-out our statistical model. Our exploratory data analysis followed the following work flow: variable identification, univariate analysis, missing values treatment, outlier treatment and variable transformation.

Using R we used barplots to study discrete variables like non-profit organizations (NPOs) type, number of employees, and geographic spread; and we used density plots to study continuous variables like revenue and assets. The graphs confirmed that the variability of the nonprofit sector with respect to size and revenue is large. Thus, to obtain a more nuanced evaluation of how different sized NPOs operate, we disaggregated our data into 2 types of categories, each with 3 classes: small (<100), medium (100<...<1000) and large (>1000) nº of employees and small(<1billion), medium (1<...<100 billion) and large (>100 billion) revenue in US dollars.

Our univariate analysis allowed us to validate an initial hypothesis: although education and health NPOs make-up only a tiny fraction of total NPOs in the sector, these NPOs make up the bulk of high-revenue within the sector. Given the magnitude of these two types of NPOs, we decided to run two analyses—including and excluding these types of NPOs—in our model in order to compare and contrast the further nuance our analysis of the sector. This analysis will be incorporated into the descriptive statistic section of the web application.

We looked for missing values using the is.na() function and found that our dataset contained no missing values. We then treated outliers by first locating them with a function and for-loop and then by grouping them seperately in histograms and density plots. We found that outliers increased the error variances and decreased normality amongst our variables.

Given the large amount of variables in our dataset, we relied on computational methodologies to select the variables that would predict the total revenue target variable. Using the glmnet package in R, we implemented the Lasso regression method to achieve a sparse solution through penalization. By implementing this methodology, the Lasso model enabled us to train and test our model.

To ensure that the unimportant coefficients shrunk properly, prior to implementing Lasso, we transformed our variables by standardizing the data. We then established training and validation sets: the training set consisted of 70% of the total data and the validation set consisted of the remaining 30% of data. Using glmnet we provided a sample of penalizing elements for the Lasso model, from which the tuning parameters that minimizes MSE was chosen. By fitting training x variables into the model we predicted outcomes in validation set—total revenue—to be compared with total revenue stored in the validation set, such that MSE could be computed. If the training model produces reasonable MSE based on the validation set, then the model (and so the selector variables) were adopted.