Predicting and Exploring Nonprofit Employment Using Demographic Data

Robert Ortiz

DATA 72100 Advanced Statistical Methods

Professor Johanna Devaney

Introduction

Throughout the semester we explored how to work with supervised and unsupervised machine learning. The course consisted of cleaning, training and testing sets, and applying different machine learning methods to understand patterns and make predictions. Learning these methods were useful and the learning curve became more manageable by the time we reached clustering I found supervised models to be more difficult to perform which was tied to my imbalance dataset. I relied heavily on online documentation and YouTube (even some TikTok's) to better digest the course material.

Dataset Description and Research Goal

For this project, I opted to use data from the U.S. Census Bureau Public Use Microdata Sample (PUMS). The dataset contains individual-level demographic and employment information for people living in the United States. I chose this dataset because it is publicly available and relatively related to my research goal about employment and workforce patterns in the nonprofit sector. Making the determination to use PUMS data was uncomplicated partly because I was already using it for another project. While I thought converging two projects was strategically sound, it ended up not being the best approach. There was a distinct disconnect due to the files that were appropriate for each project. For this assignment the API connection that I worked with was not useful and led to duplicative work. The API connection aggregated the data which was detrimental when attempting to run the analysis I was looking to complete.

The main variable I decided to focused on was whether a person works in the nonprofit sector, which was constructed using the Census variable COW (Class of Worker), where individuals were labeled as nonprofit workers (COW=2). Using class of worker created a binary target

variable labeled nonprofit_worker, where 1 indicates nonprofit employment and 0 indicates all other employment (non-nonprofits (government, private, etc.)). Initially I included all ages, however, after some reflection I decided to focus on adults (age 18 and older). In retrospect I should have probably defined the adult as civilians who are 21 years of age or older (I have a couple of reasons that I will not get into in this paper). After filtering the data to include only adults (age 18 and older) who were employed, the dataset showed that nonprofit workers made up only about 10 – 11% of the working population (data sample). This strong class imbalance became a central challenge throughout the analysis. To stray away just a bit from academic language and rely on somewhat metaphorical phrasing, it was *a thorn I couldn't get away from.*

Preparation and Data Cleaning

The initial dataset was aspirational and included the variables age (AGE), gender (SEX), education level (SCHL), employment status (ESR), class of worker (COW), disability status (DIS), and region (which was constructed from state codes). Cleaning the data consisted of isolating the target variable and dropping any rows with missing values. Apart from region, the process was not overly complicated. The layer of complexity added by region was tied to API that was initially used. The dataset flowing through the API had region as a variable.

After cleaning the data and troubleshooting through the construction of the region variable the data was split into training and testing sets. The sets were constructed using a 75/25 split, with stratification to preserve the nonprofit vs. non-nonprofit class balance. The resulting class proportions were the same in both sets, with nonprofit workers representing roughly one-tenth of observations. The results were not that different from the descriptive statistics that I obtained when running the API version of the initial run through.

Missing values were handled using median imputation for numeric variables (AGE and SCHL) and most-frequent imputation for categorical variables (SEX, DIS, and REGION). Using the approach ensured that no observations were lost (rows dropped) due to missing data. Finally, categorical variables were converted using one-hot encoding to dummy variables, and numeric variables were standardized/scaled when required by the models.

Analysis via Data Visualization

Several visualizations were used to interpret and understand the data. Histograms of AGE and SCHL showed that both variables were skewed rather than normally distributed. Scatter matrix plots allowed for the conclusion that there were no strong linear relationships between age and education. Bar plots of demographic groups revealed some consistent patterns such as women having a higher employment rate than men (which has been confirmed by numerous publications). Another pattern revealed is that nonprofit employment increased with educational attainment, particularly for individuals at higher levels of education (graduate and professional degrees, etc.). Interestingly when looking at disability it showed very small differences in employment rates. These visual patterns led to the conclusion that education and sex were likely to the strongest predictors of nonprofit employment.

Supervised Learning

The two supervised learning models used were Logistic Regression (LR) and Random Forest (RF). These models were chosen because they represent a linear classifier and a flexible, non-linear ensemble method. Analyzing the data in the two different approaches provides a more holistic analytical approach.

Default Models (Baseline)

The LR and RF models were initially processed using default parameters. The default models served as the baseline models used to compared with the weighted models. As expected due to the prevalent class imbalance, the models performed poorly at identifying nonprofit workers. Although accuracy appeared to be high, precision, recall, and F1-scores for the nonprofit variable were nearly zero. This occurred because the models mostly predicted the majority class (non-nonprofit). Cross-validation (CV) results confirmed the issue. Hyperparameter tuning using grid search improved results slightly but did not solve the imbalance problem.

Balanced Models (Class-Weighted)

To address this issue, class weights were added to both models which are referred as the "balanced" models. Applying class weights significantly improved performance: balanced Logistic Regression (BLR) achieve a recall of 0.635, while balanced Random Forest (BRF) achieved a recall of 0.666 and among all the models the highest F1-score.

Receiver Operating Characteristic (ROC) curve analysis showed that both models can tell the difference between nonprofit and non-nonprofit workers (moderate discrimination ability), with AUC values of approximately 0.67 for Logistic Regression and 0.68 for Random Forest. The Precision–Recall curves show that even though the models found more nonprofit workers (higher recall), they also made many incorrect nonprofit predictions because the two groups overlap a lot.

Feature importance analysis showed that features aligned with education were among the strongest predictors. Sex, particularly workers who identified as female, increased the likelihood of being employed in the nonprofit sector. Finally, disability did not display a high level of predictive import. Overall, the BRF model performed best for predicting nonprofit employment.

Principal Component Analysis for Feature Selection

Principal Component Analysis (PCA) using Truncated SVD was applied to the training data to reduce the number of features while keeping most of the information in the data. The used of k components was necessary to retain 95% of the variance in the data.

When PCA was applied before the RF model, F1-scores showed only a very small change compared to the baseline model without PCA. The difference in mean F1-score was minimal, indicating that PCA did not meaningfully improve supervised model performance. This suggests that RF already handled the high-dimensional feature space effectively.

Unsupervised Learning and Clustering

Clustering algorithms were applied without using the nonprofit label to run unsupervised models. Three clustering methods were tested using only the training data (as per the assignment): k-Means, Agglomerative Clustering (AC), and DBSCAN. Each method was evaluated with and without PCA using both Adjusted Rand Index (ARI) and Silhouette Score (SS). The results showed that k-Means w/o PCA achieved the highest SS (~ 0.33). AC and DBSCAN produced slightly weaker results. DBSCAN struggles to form meaningful clusters, and many times labeled points as noise. ARI values across all methos were close to 0, which indicates that the clusters did not align well with nonprofit worker/employment labels. PCA slightly improved DBSCAN results but did not change the overall conclusion. Two of the PCA plots showed that nonprofit and non-nonprofit workers showed heavy overlap, which helps explain why the clustering models did not work well.

Conclusion and Reflection

If I were to redo this project, I would consider adding variables such as income, or occupation to improve both prediction and clustering. I would also explore resampling techniques and threshold tuning to further improve minority-class prediction. More importantly, I would be more thoughtful about the data required to perform analysis for different classes.

The effect of using different data connectors (API vs. Excel file) led to delays in identifying the issues and constant breaks in the code. Compared to other courses, the file used in this project was massive and led to breakage in Google Colab which required at times unnecessary recalling of packages. The duplicative coding caused some confusion in the sequence which improved only after visuals were loaded.

Overall, this project, in particular the dataset I opted to use, demonstrated the importance of careful data preparation, visualization, and evaluation when working with real-world, imbalanced datasets.