**Capstone Project Summary Report**

Imputation of Industry and Occupation Categories

Marin Lolic, Casey Nguyen, Ahmed Soliman

Introduction

The goal of this capstone is to provide the BLS with completed data for industry and occupation across different decades (1970s to present). There are two primary resources. The first is dual-coded files, which have a very large number of samples, but are only available for a few of the time periods. The second is crosswalks, which can collectively cover all the time periods, but have far fewer samples.

The team that worked on this project last year created 240 different random forest models to predict every combination of industry and occupation categories, mostly utilizing two “universal crosswalks” that they stitched together from existing crosswalks. Our idea was to greatly simplify this process, creating two neural network-based models (one for industry, one for occupation), hopefully improving accuracy in the process.

Model Construction Process

Our first model was for industry, but the occupation model follows the same steps. We started with the universal crosswalk, which contained mappings for all 7 periods (1971, 1983, 1992, 2003, 2009, 2014, 2020). We then randomly masked out features in each row to create our input dataset (the full crosswalk is the target output). Next, we turned both the input dataset and the output into one-hot encodings since the industry categories are discrete variables. We split our dataset into three: 80% for training, 10% for validation, and 10% for test. Finally, we built the neural network, which we diagram below (Figure 1). The network begins by condensing each of the period input matrices into smaller matrices, largely to manage the total number of parameters. These condensed representations are then concatenated before being fed through a single fully connected layer. We add a dropout layer to mitigate overfitting, then map to 7 output matrices (one for each time period).

We trained the network for however many epochs it took to start overfitting or for performance improvements to stop (generally 3-5 epochs). We then checked accuracy on the validation dataset and used the feedback to tune our hyperparameters (optimizer, activation function, number of units per layer, number of layers, use of dropout). Finally, when we obtained our best model on the validation dataset, we ran it on the test set to calculate its performance. Overall performance was strong, but varied significantly by time period and between the industry model and the occupation model (see Figures 2 and 3).

Extending the Project

Our sponsor had an additional request - to see how our methodology would work using a dual-coded data set. We used the Treiman dataset, which maps both industry and occupation for approximately 125,000 individuals between 1970 and 1980. We created a very similar network to our crosswalk models, though it had some differences. First, it considered both industry and occupation at the same time, since that data was available. Second, it only considered two periods (1970 and 1980) since those were the only ones available in the Treiman data. Finally, there was no need to mask, since we could use the 1970 data as the input and the 1980 data as the target output. The holdout set performance of this dual-coded model was even better than for the crosswalk model, likely due to the large sample size and the availability of both industry and occupation data at the same time (see Figure 4).

We then looked at a dataset from 1977 (coded using 1970 variables), and used it to impute the 1980 industry and occupation data. We found a very close correspondence between the distribution of our imputed data and the distribution of actual 1980 data (see Figures 5-6). This was true for both industry and occupation. While this does not guarantee that our imputation methodology is correct, it suggests that we obtained reasonable results.

Conclusions and Further Research

There are several noteworthy results from our research. First and foremost, it is possible to use neural networks to obtain high-quality results in categorical imputation problems. Moreover, neural networks can do so far more efficiently (in terms of models trained) than many competing methods. Our work also demonstrates that different types of data can be used for this type of imputation, though as in almost all forms of machine learning, more data is better. This brings us to our first recommendation for future research - the use of as many dual-coded data sets as possible to train imputation models. Ideally, future researchers would have one or more dual-coded data sets to cover all the relevant time periods, much as the crosswalks do currently. Our second idea for future research is to investigate using embeddings instead of one-hot encodings for industry and occupation categories. The use of embeddings has become popular in natural language processing to represent words, and a similar concept could be used to transform industry and occupation categories into dense vectors.

Figure 1: Industry Model Architecture

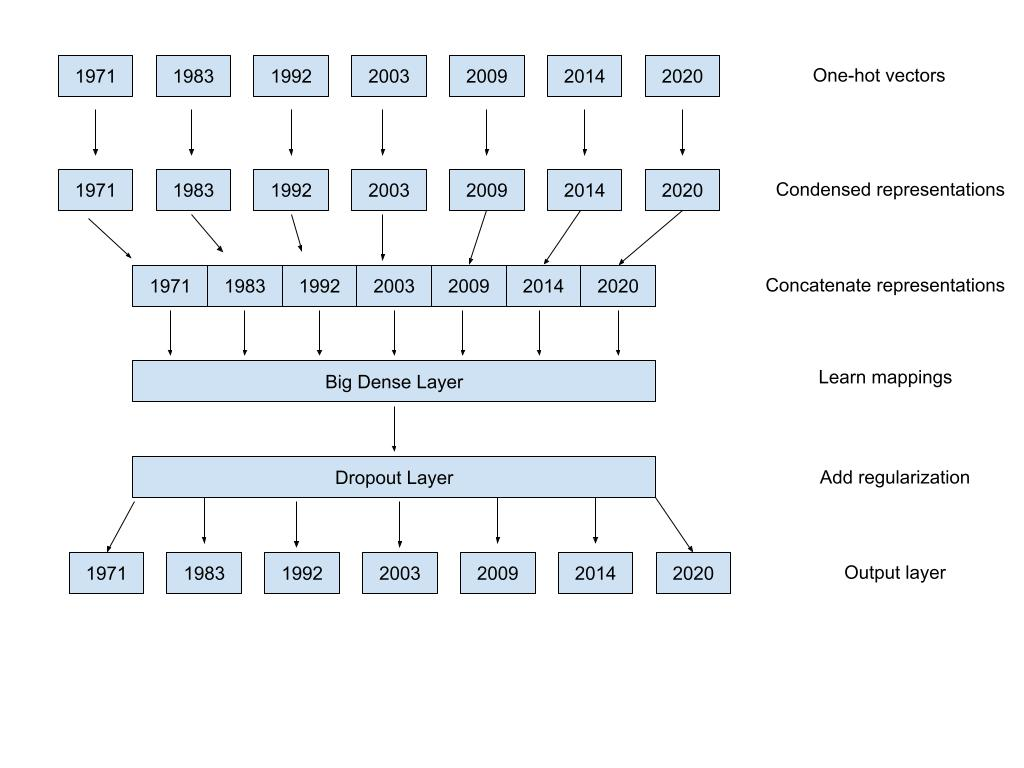


Figure 2: Industry Model Test Performance

|  | **Top 1** | **Top 5** |
| --- | --- | --- |
| 1971 Accuracy | 72.9% | 93.8% |
| 1983 Accuracy | 86.5% | 96.7% |
| 1992 Accuracy | 87.5% | 97.7% |
| 2003 Accuracy | 91.3% | 96.2% |
| 2009 Accuracy | 91.5% | 96.0% |
| 2014 Accuracy | 91.5% | 96.4% |
| 2020 Accuracy | 89.3% | 96.1% |
|  |  |  |
| **Average** | **87.2%** | **96.1%** |

Figure 3: Occupation Model Test Performance

|  | **Top 1** | **Top 5** |
| --- | --- | --- |
| 1972 Accuracy | 59.5% | 78.3% |
| 1983 Accuracy | 90.6% | 97.4% |
| 1992 Accuracy | 89.2% | 97.2% |
| 1995 Accuracy | 89.4% | 97.4% |
| 2003 Accuracy | 84.9% | 91.4% |
| 2011 Accuracy | 83.5% | 90.5% |
| 2020 Accuracy | 78.2% | 88.9% |
|  |  |  |
| **Average** | **82.2%** | **91.6%** |

Figure 4: Treiman Data Model Test Performance

|  | **Top 1** | **Top 5** |
| --- | --- | --- |
| 1980 Industry Accuracy | 94.1% | 99.7% |
| 1980 Occupation Accuracy | 85.7% | 98.0% |
|  |  |  |
| **Average** | **89.9%** | **98.8%** |

Figure 5: Comparison of Industry Imputations to Real Data

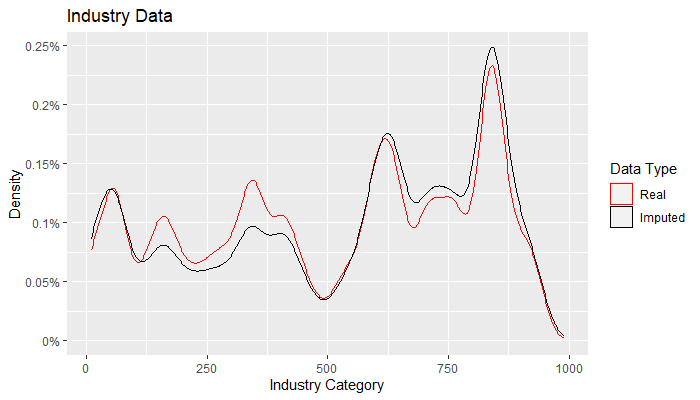


Figure 6: Comparison of Occupation Imputations to Real Data

